A Major Project Report

On

NLP Sequencing-Language Transfer Learning Models for Genomics

Submitted in the Partial Fulfillment of the Requirements

For the Award of the Degree of

Bachelor of Technology

In

CSE (Artificial Intelligence and Machine Learning)

By

D. Sarayu [21211A6614]
D. Suraj [21211A6615]
P. Rohith [21211A6638]

Under the Guidance of

Ms. Srilakshmi V **Assistant Professor**

Mrs. B Lavanya **Assistant Professor**

DEPARTMENT OF CSE(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

B V RAJU INSTITUTE OF TECHNOLOGY

(UGC Autonomous, Accredited by NBA & NAAC)

Vishnupur, Narsapur, Medak, Telangana State, India – 502 313

2024-2025

B V RAJU INSTITUTE OF TECHNOLOGY

(UGC Autonomous, Accredited by NBA & NAAC)
Vishnupur, Narsapur, Medak, Telangana State, India – 502 313

DEPARTMENT OF CSE (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

CERTIFICATE

This is to Certify that the Major Project Entitled "**NLP Sequencing-Language Transfer Learning Models for Genomics**" Being Submitted By

D. Sarayu	[21211A6614]
D. Suraj	[21211A6615]
P. Rohith	[21211A6638]

In Partial Fulfillment of the Requirements for the Award of Degree of Bachelor of Technology in CSE(Artificial Intelligence and Machine Learning) to B V Raju Institute of Technology is Record of Bonafide Work Carried Out During the Period From December 2024 to April 2025 by Them Under the Supervision of

Ms. Srilakshmi V **Assistant Professor**

Mrs. B Lavanya **Assistant Professor**

This is to Certify that the Above Statement Made by the Students are Correct to the Best of Our Knowledge.

Ms. Srilakshmi V
Assistant Professor

Mrs. B Lavanya
Assistant Professor

The Major Project Viva-Voce For This Team Has Been Held on _____.

External Examiner

Dr. G Uday Kiran
Program Coordinator

B V RAJU INSTITUTE OF TECHNOLOGY
(UGC Autonomous, Accredited by NBA & NAAC)
Vishnupur, Narsapur, Medak, Telangana State, India – 502 313

CANDIDATE'S DECLARATION

We Hereby Certify that the Work Which is Being Presented in the Major Project Entitled "NLP Sequencing-Language Transfer Learning Models for Genomics" in Partial Fulfillment of the Requirements For the Award of Degree of Bachelor of Technology and Submitted in the Department of CSE(Artificial Intelligence and Machine Learning), B V Raju Institute of Technology, Narsapur, is an Authentic Record of Our Own Work Carried Out During the Period From December 2024 to April 2025, Under the Supervision of Mrs. B Lavanya, Assistant Professor and Ms. Srilakshmi V, Assistant Professor.

The Work Presented in this Major Project Report Has Not Been Submitted By Us For the Award of Any Other Degree of This or Any Other Institute/University.

D. Sarayu	[21211A6614]
D. Suraj	[21211A6615]
P. Rohith	[21211A6638]

ACKNOWLEDGEMENT

We stand at the culmination of a significant journey, one that has been both challenging and rewarding. The success of our major project is not solely a reflection of our efforts but a testament to the invaluable support and guidance we have received from many quarters. It is with deep gratitude that we acknowledge those who have made this achievement possible.

Foremost, we extend our sincerest appreciation to Mrs. B Lavanya, Co-supervisor, and Ms. Srilakshmi V, Supervisor, whose expertise and insightful supervision have been pivotal in navigating the complexities of this project. Their unwavering support and encouragement have been our guiding light throughout this journey.

Special thanks are due to Ms. Srilakshmi V, Project Coordinator, whose assistance and guidance have been instrumental in the successful execution of our project. Her dedication and support have been a source of inspiration and motivation.

We reserve our utmost gratitude for Dr. G Uday Kiran, Program Coordinator of the Department of CSE (Artificial Intelligence and Machine Learning), whose leadership and academic guidance have enriched our learning experience and contributed significantly to our project's success.

Our journey would not have been the same without the constant encouragement, support, and guidance from the esteemed faculty of the Department of CSE (Artificial Intelligence and Machine Learning). We are deeply thankful to everyone who contributed to our journey, whose belief, guidance, and support have been crucial to our achievement. This project reflects not only our academic efforts but also the collaborative spirit and collective wisdom that guided us.

D. Sarayu [21211A6614]
D. Suraj [21211A6615]
P. Rohith [21211A6638]

ABSTRACT

Recent trends in NLP, particularly those associated with transfer learning, have created a new area of possibilities for analyzing genomic sequences. This work will analyze the comparative performance of DNABERT versus a joint BERT-LSTM solution in identifying mutations within viral genomes, using data from the NCBI Viruses database. While DNABERT, which is informed by genome-derived k-mers, has its strengths in capturing contextual features of sequence data, LSTM is capable of compensating for this by modeling long-range dependencies. Different evaluation metrics like accuracy, F-score, and generalization ability were used to assess the performance of the model in mutation-classification tasks, revealing that NLP-based deep learning models are superior to classical methods. Not only does DNABERT-2-LSTM integration predict with improved accuracy, but a reliable framework for genomic analysis is also obtained.

This study offers the advantages of transformer-based models for understanding variation within viral genomes, with future work focusing on improving model design and increasing dataset size for wider applications. Beyond the metrics, this study highlights the real-world significance of applying NLP techniques to genomic data. The hybrid DNABERT-LSTM model demonstrated not only stronger performance but also greater adaptability to diverse viral mutation patterns. By bridging deep learning with biological insight, the approach offers a step toward more responsive tools for monitoring viral evolution. Such integration is not just a technical innovation but a potential asset in public health preparedness. As the world continues to face viral threats, tools like these can help accelerate our understanding, response, and resilience. This work ultimately contributes to a growing movement that blends AI with life sciences to address challenges with global implications

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABBREVIATION	FULL FORM
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CAGE	Cap Analysis Gene Expression
CNN	Convolutional Neural Networks
DL	Deep Learning
DNA	Deoxyribonucleic acid
LSTM	Long Short- Term Memory
ML	Machine Learning
NLP	Natural Language Processing
NCBI	National Center for Biotechnology Information
RNA	Ribonucleic acid
SARS-Cov-2	Severe Acute Respiratory Syndrome Coronavirus
TFBS	Transcription Factor Binding Sites

CHAPTER 1
INTRODUCTION

This section describes the application of Natural Language Processing (NLP) models in genomic mutation detection and classification, highlighting their potential to transform genomic analysis. It covers the advancements brought by models like DNABERT-2 and Long Short-Term Memory (LSTM) networks, as well as the role of transfer learning in overcoming challenges in genomic sequence analysis.

Natural Language Processing (NLP) has become a powerful tool in genomics, treating DNA sequences like language to predict mutations. Models like DNABERT-2 and Long Short-Term Memory (LSTM) networks outperform traditional alignment-based approaches by capturing both short and long-range dependencies in DNA using deep learning. DNABERT-2 uses k-mer tokenization and bidirectional attention to detect critical mutation types. When combined with LSTM, which remembers sequence patterns over long distances, the hybrid model improves classification accuracy on viral datasets such as those from NCBI. This approach enables better tracking of mutation-driven pathogen evolution and supports large-scale mutation studies.

Genomic sequence classification has wide-ranging applications, especially in clinical diagnostics, personalized medicine, epidemiological surveillance, and drug resistance profiling. In clinical settings, identifying mutations is crucial for diagnosing genetic disorders, tailoring treatments, and predicting responses to medication. For infectious diseases, mutation classification helps identify new pathogen strains, track outbreaks, and adapt treatment strategies accordingly. Despite the advancements, challenges like class imbalance (where some mutations are underrepresented), sequencing errors, and computational resource requirements persist.

1.1 NLP FOR GENOMIC CLASSIFICATION

Natural Language Processing is a transformational tool in genomics. The most realistic scenario for linguistics modeling is to apply it to DNA sequences. With the classifying mutations in genomic sequences, it has become an integral part of modern computational biology in viral evolution, disease mechanism determination, and therapeutic development. The traditional bioinformatics methods, primarily alignment-based ones, can hardly deal with complicated patterns, structural variants, or highly variable pathogens. Incorporating deep learning into this field entailed a host of brighter, larger-scaled, and more understood approaches such as these: NLP-inspired models like DNABERT-2, which is an offshoot of BERT for DNA sequences, and Long Short-Term Memory (LSTM) networks. DNABERT-2 employs k-mer tokenization to break down DNA sequences into contiguous overlapping sub-sequences, presenting essential local motifs and global dependencies as a basis for the mutation analysis.

Its bidirectional attention mechanism allows for context-sensitive comprehension, making it able to tell the difference even between subtle forms of mutation such as synonymous and non-synonymous substitutions. LSTM networks, besides this transformer-based method, can also be used to do a sequential modeling with the capacity to maintain long-range dependencies in genomic data. Even increasing the representation space with DNABERT-2 embeddings, LSTMs strengthen this combined framework because their memory-based learning enhances classification tasks. Use of the combined framework capitalizes on the strengths of both architectures making it suitable for characterizing and classifying mutations in large datasets of viral sequences. Beyond this are the availability of curated resources such as the NCBI Viruses dataset. This repository comprises annotated viral genomes, labelling mutations, thus yielding a rich and broad training and validation ground for models on virus classification. Its application

guarantees robustness and generalizability to a variety of virus families and strains.

Development of genomic language models such as DNABERT, DNABERT-2, and more recently GenomeGPT and GenFormer reflect a new trend where NLP architectures are adapted into biological data. They symbolize a break from the past tradition and indicate the far-reaching capability of language modeling in the analysis of complex unstructured large-scale DNA sequences.

This trend of converging computational linguistics with biology shows that the NLP efforts on genomics have great prospects when it comes to detecting and classifying mutations or putting them in context. This would immediately accelerate research in virology and personalized medicine while laying a general groundwork for real-time AI-bio discovery.

1.2 MOTIVATION

This project is motivated mainly by an urgent need for rapid and scalable precision tools that classify mutations in genomic sequences during viral outbreaks like COVID-19. Due to the high mutation rates of viruses, such as SARS-CoV-2, it is also important to carry out real-time detection and classification of mutations to meet public health measures, vaccine development, and treatment strategies. Existing bio-informatics tools generally tend to fall short in terms of flexibility and efficiency to respond to the rapidly evolving viral genomes.The use of advanced deep learning models such as DNABERT-2 and LSTM presents some advantages: one benefit is that pre-trained embeddings derived from DNABERT-2 allow good generalization across various genomic sequences, especially if labeled data are limited. Second, LSTMs are capable of capturing long-range dependencies and can so detect more complex mutation patterns which may have otherwise been missed by classic approaches. These deep-learning models could greatly enhance the speed and accuracy of classifying mutations within

the workflows, thereby offering insights into the evolution of the viral strain and facilitating the development of better strategies for dealing with infectious diseases.

1.3 PROBLEM STATEMENT

Traditional approaches for mutation classification, such as alignment-based algorithms, are found limited in scope as they are unable to handle the vastness and intricacies of genomic data. These have dependency on **reference genomes and frequently find difficulties in subtle mutations, low-quality data, or new viral strains**. Furthermore, these new viral sequences align to the reference genome, which is often misleading in the case of their occurrence of rare or unknown mutations.

Another obstacle is the dependency on large, labeled datasets. Annotation of genomic sequences is an expensive and time-consuming task, which adds difficulty in developing solid machine learning models. This problem becomes more pronounced with very little labeled data available concerning rare or newly defined mutations, as seen in the case of SARS-CoV-2 and other rapidly evolving viruses.

The current project focuses on addressing these issues using both DNABERT-2 and LSTM networks. Specifically, DNABERT-2 capitalizes on the idea of transfer learning as well as modeling sequential dependencies and enables a more efficient classification of mutations without the need for vast amounts of labeled data. LSTM models, on the other hand, are improved in their ability to capture long-range dependencies and deliver a highly scalable and efficient mutation classification approach even with limited data and complex sequence variations.

1.4 OBJECTIVE

- The objective of the project is to create a mutation classification framework based on deep learning using DNABERT-2 and LSTM to ascertain point mutations, insertion, deletion, and structural variants in genomic sequences.
- Transfer learning with DNABERT-2 would grasp both local and global sequence relationships to enhance classification accuracy in particular to the case of limited labeled data. Integrate an LSTM to describe long-range dependencies in DNA sequences, as are necessary for observing mutations with effects over larger genomic sections.
- Permit analysis across a diverse set of viral genomic datasets, including those from the NCBI Viruses repository, to homologate mutation classification across multiple viral species.

1.5 MUTATIONS AND TYPES

These changes in the DNA sequence of the genome of an organism are called mutations. The alterations can be the slight difference of even a single pair of nucleotide bases to significant chromosomal rearrangements. It is a major source of genetic diversity in evolution and also increases the variation among individuals. Mutations occur spontaneously or may be caused by external factors such as radiation or chemicals. They do not always cause harm; some mutations create genetic disorders or diseases, while others are neutral or beneficial to the individual in terms of offering advantages. The understanding of mutations is key to genetics, evolution, and the mechanisms of diseases.

- **Point Mutations (Base Substitutions)**

  Point mutations (base substitutions) are some common types of genetic mutations found in the host and viral genome. These mutations change one nucleotide base that is capable of changing the function or expression of genes. Point mutations may be

classified as silent mutations when the nucleotide change does not change the amino acid coded for because of redundancy in the genetic code. Silent mutations do not change the structure of the protein, but they may interfere with regulatory mechanisms, including mRNA stability and splicing.

In a missense mutation, the alteration of the nucleotide sequence results in the incorruption of a different amino acid into the protein. Depending on the specific substitution and where it occurs, this change may have no effect or may be very serious, causing the protein to alter signifi cantly in structure or function. Nonsense mutation is the most disruptive type because it substitutes a stop codon for a codon in the mRNA, terminating protein synthesis, usually resulting in a loss of function, and is implicated in serious genetic disorders or impairments in viral protein activity.

- **Indels- Insertions and Deletions**

  They are generally known as insertions and deletions or, more succinctly, as indels: either the addition by insertion of nucleotide bases to DNA or the deletion of nucleotide bases from DNA. They come in varying sizes and make a considerable difference in the protein-coding regions involved. If the added or deleted bases are not a multiple of three, the frame used to read the gene changes- frameshift mutation. This change usually results from changes in downstream codon alignments, frequently resulting in a radically altered or nonfunctional protein.

  Frameshift mutations most often produce early stop codons and, consequently, truncated proteins. In compact viral genomes, these modifications may affect replication efficiency, recognition by hosts, or mechanisms for evasion of an immune response. Indels within non-coding or regulatory regions can alter the levels or timing of gene expression. Identifying and classifying these

variations leads to better knowledge of their function in pathogenicity or treatment resistance.

- **Structural Variations**

  The alterations of DNA involve structural variations; they are changes that affect long parts of DNA and range from hundreds to millions of base pairs in length. Duplications, inversions, and translocations are the three types of structural variations, each having a specific effect on the structure and function of the genome. Repeated segments of DNA result from duplications and thus could even lead to changes in gene expression or create new gene variants. Duplications in viral genomes may affect virulence or spread because of altered dosage of the genes.

  Inversions are interested when a fragment of DNA is flipped within the chromosome. Any critical gene or regulatory element disrupted might hence interfere with function or expression, particularly in tightly regulated viral systems. The movements of DNA segments between non-homologous chromosomes or different locations of the same chromosome are called translocations. These rearrangements cause gene fusions or misregulated gene expression; they are also commonly associated with complex diseases and cancers.

  Though less frequent than point mutations and insertions/deletions, structural variations have biological significance and are rapidly becoming a niche area in genome-wide mutation-classifying studies.

1.6 PLAN OF IMPLEMENTATION

The mutational classification paradigm uses Natural Language Processing and deep learning methodologies for viral genomic sequence analysis. The entire course of action can be expressed in the following steps:

- **Datasets Acquisition:** All genomic sequences were downloaded from the NCBI Viruses database, which carries multiple viral genome samples along with mutation information. Each sequence has its classification tagged as either Wildtype (0) or Mutated (1), intending to create a binary classification. The pro-dataset is an asset both for the training and evaluation stages because it contains unparalleled mutation across viral families.
- **K-mer Counting and Sequence Preparation:** Generally, MUOT extracts from a DNA sequence superimposed sub-sequences by k-mer tokenization: ATGCTG-TGCTGA-GCTGAT are for example 6-mers. This is very important pre-processing to linear genomic data suitable for transformer-based tokenization.
- **Tokenizing using DNABERT-2:** The k-mer sequences will be fed into a pretrained model DNABERT-2; this serves as a genomic language encoder. DNABERT-2 is specifically made for DNA sequencing, as it will be able to produce contextual embeddings modeling local motifs and overall attributes of the sequence by means of bidirectional transformer attention.
- **Customizing Transformer Inputs:** The output results obtained from tokenization will be harmonized to meet the transformer model requirements such as position encoding and input embedding layers so that they correspond to the DNABERT-2 transformer sections while going forward.
- **Feature Extraction with DNABERT-2 Layers:** The tokenized updated strings gradually enhance their representation by various layers of transformer blocks. These layers from an input sample comprise multiple attention heads along with feedforward layers, thus yielding very rich semantic embeddings that take into account the context of the sequence.
- **Sequential Modeling with LSTM:** The DNABERT-2 embeddings are fed into an LSTM (Long Short-Term Memory) Network. An LSTM is good at learning long-term dependencies in sequences

which enhance the prediction by simulating the sequential flow and transition of context in genomic data examples.

- **Model Output – Binary Classification:** Final output from the LSTM serves to classify the mutation class. The system is designed to carry out binary classification according to these two types and will classify any given sequence into only one of two: Wildtype (0) or Mutated (1).
- **Model Training and Fine Tuning:** The model was developed using pre-trained DNABERT-2 embeddings and subsequently fine-tuning using labelled viral genomic sequences sourced from NCBI Viruses. Only the LSTM layers and the classification head were optimised for rapid adaptation because class imbalance was mitigated as follows: stratified data splits, early stopping, and class-weighted loss during training.
- **Model Evaluation:** The assessment of the model was carried out on three different datasets of viruses, namely, Human Virus, Influenza Virus, and SARS-CoV-2, under fixed test conditions using accuracy and F1-score. The amount of the performance that was measured would be in respect of how effective is mutation classification. It was found that human virus produced the highest accuracy that was found to be 80.0%, while the lowest accuracy was observed in the case of the Influenza virus, which was found to be 76.7%.
- **Final Testing and Validation:** LSTM combined with DNABERT-2 gives very promising results in terms of accuracy and F1-scores for the three datasets, highlighting the extended applicability of the model to a variety of viral genome structures. Regular performance indicates the capability of the model to identify subtle drastic patterns in the mutation sequences. One more thing that clearly substantiates the possibility of applying NLP-based deep learning approaches in scalable and accurate mutation classification is the same.

1.7 ORGANIZATION OF THE WORK

Chapter 1 introduces the intersection of deep learning, NLP, and genomics, highlighting the importance of mutation classification in viral DNA and the rationale for using language models like transformers in genomic analysis. Chapter 2 reviews related work, tracing the evolution from traditional ML to deep learning, focusing on DNABERT and its hybrid use with LSTM for contextualized DNA sequence classification. Chapter 3 details the proposed DNABERT-2 + LSTM model, covering data preprocessing, k-mer tokenization, embedding, and sequential learning using viral genome data from NCBI. Chapter 4 presents the experimental setup, evaluation using accuracy and F1-score across three viral datasets, and a performance comparison with baseline models to validate the hybrid approach. Chapter 5 concludes the study, summarizes key findings, and outlines future directions such as hyperparameter tuning, dataset expansion, and integration of multi-modal biological data.

CHAPTER 2

REVIEW OF LITERATURE

This chapter presents a comprehensive review of existing literature related to the research topic. It discusses previous studies, theories, and models that form the foundation of the current work. By analyzing relevant research papers and publications, the chapter identifies key findings, trends, and gaps in knowledge that the study aims to address. It provides the necessary background and justification for the chosen research direction.

2.1 INTRODUCTION

Natural Language Processing (NLP) and genomic sequence analysis together have made tremendous strides toward mutation classification. On the one hand, conventional approaches such as sequence alignment and rule-based techniques provide reliable means for mutation classification and, on the other hand, they are generally unable to grasp the complex context and long-range dependencies that exist in such genomic sequences. Instead, researchers are increasingly turning toward a variety of deep learning approaches, especially transfer learning, in the hope of enhancing classification performance on complex biological datasets. Transformer-based models, particularly DNABERT-2, have seemingly great promise in trusting handy genomic representations, applying the principles of BERT to DNA sequences via k-mer tokenization. With pretraining on larger DNA datasets, DNABERT-2 very well captures both sequence patterns and contextual relationships that serve in detecting mutations. But although much attention accrues to the sequence information, at times transformers remain deficient in modeling the very long sequences of DNA.

To overcome these limitations, hybrid types of innumerable investigations that marry the contextual capabilities of the transformers with the temporal modeling strengths of recurrent networks have been

developed recently. Long short-term memory (LSTM) networks, a type of recurrent neural network, are particularly useful because they can maintain information for long periods of time across long sequences. The ability of the LSTM, coupled with DNABERT-2, is to further track mutation patterns that evolve over vast areas of the genome.

This review describes the evolution of deep learning in genomics with respect to DNABERT-2 and LSTM architectures. In this section the discussion is on their application toward classifying viral mutations and how they deal with issues such as class imbalance, sparse labeling, and widespread mutation diversity. In particular, the review focuses on important trends that exploit pre-trained genomic embeddings to boost mutation detection performance against multiple viral families, Human Virus, Influenza, and SARS-CoV-2.

2.2 REVIEW ON DNABERT IN GENOMICS ANALYSIS

Zhou et al. [5] presented DNABERT-S, an extension of the DNABERT model that incorporates species-aware embeddings for genomic sequence classification. The model uses a novel Manifold Instance Mixup (MI-Mix) approach and a Curriculum Contrastive Learning strategy to enhance its understanding of species-specific sequence variations. It demonstrates superior performance in species differentiation, especially in low-shot learning scenarios, making it ideal for metagenomic applications. This development addresses a significant gap in conventional models, which often generalize poorly across species due to lack of evolutionary context.

Fetni et al. [6] explored the application of BERT-based architectures to DNA sequence pattern recognition. Her PhD dissertation focused on adapting NLP techniques to detect functional genomic elements like motifs and repeats within DNA sequences. She demonstrated that BERT-based models can capture long-range dependencies and structural regularities that traditional methods overlook. Her research establishes the foundation for integrating

transformer-based models in bioinformatics pipelines and suggests future directions for explainability and model refinement in genomic contexts.

Kassab et al. [7] conducted a linguistic analysis of DNABERT's modeling of biological sequences, viewing DNA as a structured language with its own syntax and semantics. His doctoral work delved into how sequential dependencies are encoded in transformer models and how this information can be used for predicting gene regulatory functions. By bridging computational linguistics and genomics, Kassab highlighted the potential of BERT-based models to serve not just as predictors but also as interpretable tools for biological discovery.

Sanabria et al. [8] focused on distinguishing between word identity and sequence context in DNA language models using DNABERT. They applied interpretability tools to dissect how transformer layers and attention mechanisms treat different k-mer combinations across genomic sequences. Their results reveal that deeper layers of DNABERT capture broader functional contexts while earlier layers remain sensitive to local motifs. This insight is valuable for fine-tuning models in functional genomics tasks and enhancing their biological relevance.

Akay et al. [9] innovatively applied DNABERT to predict DNA toehold-mediated strand displacement rate constants, a key process in DNA nanotechnology and molecular computing. By training DNABERT on datasets linking sequences to kinetic parameters, they showed the model's capacity to infer dynamic chemical behaviour from static sequence information. This work expands DNABERT's utility beyond genomics into bioengineering, suggesting its applicability in designing efficient DNA-based circuits and sensors.

Wu et al. [12] explored the potential of DNABERT in cancer mutation detection. By fine-tuning the model on labeled cancer genomic datasets, they demonstrated enhanced performance in classifying

cancer-associated mutations versus benign variants. Their study underscores the potential of transformer-based models in clinical genomics and suggests DNABERT can be a reliable tool for aiding in early cancer diagnosis through high-throughput sequence screening.

Zhao et al. [14] applied BERT-based models to the prediction of CpG islands, regions of DNA with high CG content often associated with gene promoters. Their model showed high accuracy in identifying CpG islands compared to traditional statistical methods. The use of transformer embeddings allowed the model to detect subtle sequence signatures beyond simple nucleotide frequency, pointing to the utility of deep contextual learning in epigenomic annotation.

Liu et al. [15] provided a broad review of pre-trained BERT-based models applied to DNA sequence classification. They evaluated various BERT derivatives including DNABERT across tasks such as enhancer identification, mutation prediction, and splice site recognition. Their findings suggest that transformer-based pretraining offers consistent improvements across multiple genomic benchmarks, positioning these models as new standards in computational genomics.

Ren et al. [16] fine-tuned DNABERT for detecting disease-associated mutations. Presented at a machine learning in genomics workshop, their approach utilized disease-specific genomic data to adapt DNABERT's parameters for clinical variant classification. The results showed significant gains in accuracy and precision compared to baseline models, demonstrating that task-specific fine-tuning of DNABERT enhances its practical utility in genomic medicine.

Gupta et al. [17] explored transformer-based classification of mutations in oncogenes using models similar to DNABERT. Their study in Nature Communications demonstrated high accuracy in distinguishing driver mutations from passenger mutations using contextual sequence embeddings. The work reinforces the relevance of

transformer architectures in cancer genomics and paves the way for their integration into personalized medicine pipelines.

Wang et al. [18] investigated the use of self-attention mechanisms within DNABERT for identifying functional genomic elements. Their work highlighted how self-attention layers contribute to recognizing long-range dependencies and regulatory features such as enhancers, silencers, and transcription factor binding sites. Their results emphasize that interpretability of attention weights can guide biological discovery, aiding hypothesis generation in non-coding genomics.

Lee et al. [20] developed a contrastive learning framework on top of DNABERT for learning more informative genomic sequence representations. By encouraging the model to distinguish between similar and dissimilar sequence pairs, the authors improved the robustness and transferability of genomic embeddings. Their work, published in Nucleic Acids Research, contributes to the field of unsupervised learning in genomics and offers tools for better representation learning in low-label environments.

Kumar et al. [21] explored the use of BERT-based predictive modeling to assess the mutational impact on disease progression. Their model leveraged DNABERT's language understanding capabilities to identify critical mutations that influence the development of complex diseases. By incorporating domain-specific mutation profiles, the study demonstrated enhanced predictive accuracy and interpretability over traditional statistical methods. The authors reported that DNABERT was particularly effective in classifying mutations according to their potential pathogenicity. Their results highlighted the advantages of contextual embeddings in understanding genomic alterations, emphasizing BERT's superiority in genomic sequence interpretation. This approach presents promising implications for translational genomics and personalized medicine.

Shen et al. [22] utilized DNABERT for detecting genomic variants linked to neurological disorders, marking an advancement in precision diagnostics. The study employed the pre-trained transformer to capture sequence features often overlooked by conventional models. By aligning variant patterns with known disorder-associated loci, the authors achieved higher sensitivity and specificity in identifying risk markers. This research underscored DNABERT's potential in neurogenomic studies, where sequence complexity and functional annotation pose significant challenges. The model's ability to extract fine-grained information from nucleotide sequences supports its application in clinical variant interpretation and neurogenetic disease prediction.

Wang et al. [23] proposed a DNABERT-based framework for predicting genetic regulatory networks, aiming to model complex gene interactions more effectively. By treating genomic sequences as a language, the model was trained to understand the grammar of regulatory elements across multiple datasets. The framework improved predictions of gene expression and transcriptional regulation compared to previous network-based models. The integration of DNABERT allowed for accurate mapping of cis-regulatory modules and interactions between distant genomic regions. This study provides insights into the deep learning-driven reconstruction of genetic regulation, offering a scalable tool for functional genomics and systems biology.

Zhang et al. [24] developed a methodology using pre-trained DNABERT models for the functional annotation of mutations. This study focused on classifying mutations into functional categories such as benign, likely pathogenic, or pathogenic. Through extensive validation on curated datasets, the authors demonstrated that DNABERT-based annotations outperformed other models, particularly in edge cases with limited annotation evidence. The fine-tuned model provided interpretability by associating sequence features with biological impact. This work underscores the importance of transformer

models in automating genome interpretation tasks and supports the adoption of DNABERT in clinical bioinformatics pipelines.

Wang et al. [25] investigated the application of transformer-based models for enhancer and silencer identification, crucial regulatory elements in gene expression. DNABERT was trained and fine-tuned on curated enhancer-silencer datasets, demonstrating higher performance in detecting these elements compared to CNN and RNN models. The study highlighted DNABERT's ability to distinguish subtle sequence variations and structural motifs associated with gene activation or repression. Results indicated improved precision and recall, especially in non-coding regions. This research underscores the model's utility in uncovering regulatory landscapes, aiding efforts in gene therapy and functional annotation.

He et al. [27] presented a DNABERT-based model for mutation classification, targeting clinical variants of uncertain significance. The model used a transformer backbone to evaluate nucleotide context and mutation effects, producing classification outputs with high confidence levels. By training on large mutation databases, the model generalized well to rare and novel variants. The study emphasized DNABERT's contribution to genomic medicine by facilitating mutation interpretation in the absence of functional assays. The work contributes to expanding the utility of language models in the variant classification framework recommended by genomic guidelines.

Zhang et al. [28] proposed improvements to DNABERT for identifying splicing variants, which play critical roles in post-transcriptional regulation. The model was enhanced with attention mechanisms specific to exon-intron boundaries, improving detection accuracy for alternative splicing events. The authors demonstrated that their model surpassed standard models in sensitivity and variant categorization, especially for non-canonical splice sites. This research enables better understanding of transcript diversity and its role in

disease, providing tools for comprehensive splicing variant annotation using deep learning.

Zhou et al. [29] applied BERT-based deep learning techniques to identify genetic markers linked to disease susceptibility. The study focused on capturing complex sequence relationships across various disease cohorts. DNABERT's contextual analysis enabled more accurate biomarker discovery, outperforming traditional statistical methods and shallow learning models. The authors emphasized its utility in genome-wide studies, particularly for low-effect variants. This model facilitates early diagnosis and risk assessment, representing a shift toward language model-driven genomics for disease susceptibility profiling.

Wu et al. [30] explored transformer-based frameworks for DNA methylation analysis, a key epigenetic modification influencing gene expression. DNABERT was adapted to detect methylation-prone sequences, learning methylation signatures from large-scale epigenetic data. The model provided improved classification of methylation status and captured regulatory contexts missed by conventional methods. These findings demonstrate the utility of DNABERT in epigenetics research, enabling discovery of methylation patterns relevant to development, disease, and environmental responses.

Liu et al. [31] demonstrated the application of DNABERT in predicting drug-response associated genetic variants, an important step in pharmacogenomics. By analyzing sequence data from drug response studies, DNABERT accurately identified genetic markers influencing drug efficacy. The transformer-based approach improved biomarker prediction compared to traditional pharmacogenetic models. This study highlighted the model's role in advancing precision medicine by linking genetic variation to treatment outcomes. DNABERT's language modeling capability enables a nuanced interpretation of variant-function relationships in pharmacogenomics.

Chen et al. [33] introduced a DNABERT-driven model for predicting epigenetic modifications such as histone marks and chromatin states. By leveraging the contextual capabilities of transformers, the model could infer modification patterns from raw DNA sequences. Performance was validated against experimental datasets, showing robust accuracy across different epigenetic contexts. This application supports functional genomics and regulatory element discovery, emphasizing the expanding scope of DNABERT in epigenome prediction tasks.

Sun et al. [34] enhanced DNABERT performance by integrating multi-modal learning strategies for classifying disease variants. The model combined sequence data with structural and clinical annotations, improving prediction reliability. The multi-modal framework enabled comprehensive representation of variant features, outperforming unimodal baselines. This work showcases DNABERT's flexibility in handling heterogeneous biomedical data, supporting advanced variant classification in genomic diagnostics.

Guo et al. [35] fine-tuned DNABERT with domain-specific genomic datasets focused on cancer research. Their customized model outperformed baseline versions in detecting cancer-associated variants and regulatory elements. By tailoring pretraining to oncogenic sequences, the model improved classification performance and interpretability. The study emphasizes the importance of domain adaptation for transformer models in specialized biomedical contexts. DNABERT's success in oncology applications highlights its potential for broader use in disease-specific genomic research.

2.3 REVIEW ON HYBRID DNABERT UTILIZATION

Ghosh et al. [2] developed TFBS-Finder, a hybrid deep learning model that combines the contextual understanding of DNABERT with the pattern recognition strength of convolutional neural networks to

predict transcription factor binding sites (TFBS). Their approach emphasizes the complementary strengths of both architectures: while DNABERT provides a global view of sequence context, CNNs efficiently capture localized motifs. This synergy improves the model's predictive accuracy and robustness, especially when dealing with noisy genomic sequences or sequences with low conservation. The researchers trained and evaluated the model on diverse datasets and showed that TFBS-Finder outperformed existing methods in terms of precision and recall. Their work contributes to the field of gene regulation by offering a more reliable tool for TFBS discovery, which is crucial for understanding the transcriptional landscape and its implications in disease and development.

Ma et al. [3] introduced HybriDNA, a novel hybrid DNA language model that integrates the transformer architecture with Mamba2—a newer sequence modeling technique designed to handle very long-range dependencies with improved computational efficiency. HybriDNA addresses key limitations in existing transformer models such as memory inefficiency and processing constraints when analyzing long genomic sequences. By combining the best of both worlds—transformers for fine-grained contextual embeddings and Mamba2 for scalable long-sequence modeling—the model significantly improves performance on a variety of downstream tasks, including enhancer identification, chromatin state prediction, and sequence classification. The paper also discusses the importance of hybrid architectures in biological sequence modeling, especially in contexts where subtle and dispersed signals influence gene regulation over large genomic distances.

Zhang et al. [13] integrated DNABERT with Graph Neural Networks (GNNs) to model gene regulatory interactions. By embedding sequence information using DNABERT and then propagating regulatory relationships via GNNs, they created a comprehensive model for gene expression regulation. This hybrid approach capitalizes on DNABERT's

strength in sequence understanding and GNN's capacity to model complex biological networks, offering a more holistic view of transcriptional control.

Li et al. [26] introduced a hybrid architecture that combined DNABERT with CNNs to predict transcription factor binding sites (TFBSs). The model leveraged DNABERT's contextual embeddings to capture global sequence dependencies and used CNNs to refine local patterns. The combined approach showed significant improvements in both accuracy and generalizability across different TFBS datasets. The hybrid model excelled in detecting motifs with variable lengths and low conservation, often missed by traditional methods. This integration bridges the gap between sequence-level language modeling and biological motif recognition, enhancing functional genomics tools.

Li et al. [32] developed an attention-based DNABERT model tailored for genome-wide association studies (GWAS). Their model was trained to detect associations between genomic regions and complex traits using a sequence-based approach. The integration of attention mechanisms allowed for more interpretable and biologically relevant feature extraction. Results demonstrated enhanced predictive power in identifying trait-associated loci. This method improves traditional GWAS by providing deeper insight into sequence-level factors influencing phenotypic variation, offering a powerful tool for complex trait analysis.

He et al. [36] introduced a transformer-based framework tailored to single-cell genomics, a field marked by high dimensionality and noise. The model addresses the unique challenges of analyzing gene expression profiles at the single-cell level using attention mechanisms to capture dependencies across diverse cells. Their approach enables more accurate clustering and annotation of cell types, significantly improving upon traditional methods. By leveraging the contextual understanding inherent to transformers, the study demonstrated

enhanced performance in identifying rare cell populations. This work highlights the importance of deep learning in single-cell analysis and opens new avenues for understanding cellular heterogeneity. Moreover, it exemplifies the potential of language models to uncover complex patterns in high-resolution biological datasets.

Zhang et al. [37] explored DNABERT's application in virus detection and genomic analysis, where sequence specificity is critical. Their work demonstrated how DNABERT could distinguish viral genomes from host DNA with high accuracy by learning intricate k-mer patterns. The model was trained on large-scale viral genomic data, and its classification accuracy surpassed conventional tools. The study provided insights into DNABERT's ability to identify key viral motifs, which can aid in developing diagnostic tools. Their findings also suggest applicability in monitoring viral mutations and understanding viral evolution. The success of this model underscores the utility of NLP-inspired architectures in virology.

Wu et al. [38] employed unsupervised learning via DNABERT to discover rare variants, which are often missed in traditional pipelines due to low representation in datasets. By using a masked language modeling approach, they allowed the model to infer patterns and identify deviations that signify rare mutations. Their study revealed that DNABERT could generalize well across species and contexts without the need for extensive labeled data. This approach is particularly useful in exploratory studies and underrepresented populations. Their work showcases the strength of unsupervised techniques in genomic discovery and mutation profiling. It therehighlights the flexibility of DNABERT for variant detection in noisy or limited-data environments.

Xie et al. [39] proposed a novel algorithm for 4mC site recognition using a pruned pre-trained DNABERT model, integrated with artificial feature encoding. The pruning process reduced the model's complexity while retaining essential capabilities, enhancing interpretability and

computational efficiency. This hybrid model successfully captured methylation site characteristics, leading to improved prediction accuracy. The incorporation of artificial features allowed the model to better handle complex methylation patterns. Their framework represents an effective balance between model precision and efficiency, paving the way for methylation studies on resource-constrained systems. This advancement holds promise for epigenetic regulation research in both prokaryotic and eukaryotic systems.

Wang et al. [40] developed BERT-TFBS, a model leveraging transfer learning to predict transcription factor binding sites (TFBS). The model utilized DNABERT's pre-trained embeddings and fine-tuned them with TF-specific datasets, achieving superior prediction accuracy. This approach addressed the scarcity of labeled data in certain TF categories by reusing generalized DNA sequence knowledge. Their model demonstrated robustness across multiple species and binding site types. It also enabled interpretable attention maps, shedding light on binding motifs. The study underscores the value of transfer learning in genomics, where annotated datasets are often limited. Overall, BERT-TFBS marks a significant step forward in TFBS prediction.

Gupta et al. [41] presented PTFSpot, a deep co-learning model designed to predict transcription factor binding regions with universality across plant species. Integrating BERT-based DNA representations and co-attentive mechanisms, the model captured intricate TF-binding interactions. It performed impressively across diverse plant genomes, even in previously unstudied species. The co-learning strategy facilitated shared learning between TF categories, enhancing generalizability. Their model proved especially effective in low-data environments, highlighting its adaptability. This research provides a framework for plant regulatory genomics, where genomic resources are often sparse. PTFSpot represents an important contribution to cross-species regulatory prediction tools.

Li et al. [42] introduced CodonBERT, a large-scale transformer model designed for mRNA vaccine development. The model focused on codon usage bias, mRNA stability, and translation efficiency—key aspects of vaccine efficacy. CodonBERT learned biologically meaningful patterns from vast transcriptomic datasets, enabling it to suggest optimized codon sequences. The model's predictions were validated in experimental settings, showing enhanced translation in target systems. This innovation bridges deep learning and synthetic biology, particularly in the design of effective genetic constructs. Their work sets a precedent for using foundation models in vaccine design and synthetic genomics. CodonBERT exemplifies the role of AI in next-generation vaccine strategies.

Danilevicz et al. [48] explored explainable long non-coding RNA (lncRNA) identification using DNABERT in plant genomes. Their study focused on transparency and interpretability, applying gradient-based attention visualization to understand decision-making. This was critical in understanding plant regulatory networks and non-coding element functions. The DNABERT model exhibited strong performance even in novel genome assemblies. Their framework offers both predictive power and interpretability, which is rare in complex genomic tasks. This contributes significantly to plant genomics and model explainability.

Moyano Gravalos et al. [49] evaluated various NLP-based algorithms, including DNABERT, for deep learning tasks in genomics. His thesis detailed comparative analyses of BERT-style architectures across multiple biological tasks, such as classification and motif detection. The results showcased the flexibility and domain adaptability of DNABERT, especially when trained on curated k-mer embeddings. This work underscored the importance of task-specific tuning and transfer learning strategies in genomic contexts. It serves as a comprehensive guide for deploying NLP models in bioinformatics.

Zhang et al. [50] investigated BERT's performance on nucleotide sequences using non-standard pretraining regimes and various k-mer embeddings. Their analysis revealed how embedding granularity influenced downstream task performance, such as enhancer detection and gene annotation. DNABERT models with optimized k-mer settings outperformed generic configurations. The study provides critical insights into the tokenization strategies suitable for DNA data. These findings help fine-tune BERT-based architectures for specialized genomic use cases and improve their biological relevance.

Zhao et al. [51] proposed a Transformer-based deep learning model specifically tailored for DNA sequence analysis, showing enhanced performance in recognizing genomic patterns and sequence classification. Their study emphasized the power of transformer architectures in capturing complex dependencies within nucleotide sequences, outperforming conventional machine learning models in tasks such as promoter prediction and motif discovery. The results demonstrated that fine-tuned transformer models could successfully generalize across various DNA analysis tasks, indicating robust applicability in genomic research. This research underlined the shift toward deep contextual models in bioinformatics and the growing role of NLP techniques in biological data processing.

Kumar et al. [52] explored the utility of transformer models for predicting enhancer-promoter interactions, a critical aspect of gene regulation. Their study integrated attention-based mechanisms to analyze long-range genomic dependencies, allowing improved accuracy over traditional sequence-based prediction models. By leveraging large-scale genomic datasets, they trained transformer models to learn meaningful patterns and associations between distal regulatory elements and target promoters. Their findings suggest that transformer-based frameworks can capture subtle biological signals that are often missed by earlier statistical and convolutional approaches.

Sun et al. [53] combined DNABERT and hybrid CNN-LSTM architectures for promoter prediction, focusing on both sequence context and spatial dependencies. The hybrid approach utilized DNABERT for extracting semantic-rich representations and CNN-LSTM for modeling spatial features, resulting in higher prediction accuracy. Their study illustrated that integrating multiple deep learning paradigms can enhance the prediction performance for complex genomic tasks. The model demonstrated the ability to identify core promoter elements across diverse organisms, marking an advancement in transcriptional regulation modeling.

Ghosh et al. [54] developed a novel Transformer-based capsule network to predict transcription factor binding sites, introducing capsule networks into the genomic prediction landscape. The model captured hierarchical features of genomic sequences and encoded rich spatial relationships, significantly improving classification metrics over baseline models. Their framework emphasized the benefits of combining capsule networks with attention mechanisms to reflect the spatial and contextual nature of biological sequences. The results validated the feasibility of more nuanced architectures for TFBS prediction tasks.

Zhang et al. [55] introduced DNAGPT, a generalized pre-trained model designed to handle a wide range of DNA sequence analysis tasks using GPT-like architectures. This model demonstrated remarkable performance on tasks such as enhancer identification, TFBS prediction, and variant classification. By adapting generative pre-training concepts to genomic data, DNAGPT emphasized the potential of autoregressive models in learning from raw DNA sequences without task-specific fine-tuning. This versatile framework signaled a significant leap in the use of language modeling paradigms for biological sequence understanding.

Zhang et al. [56] investigated the impact of non-standard pre-training strategies and various k-mer embeddings on BERT's performance for nucleotide sequences. Their study revealed that k-mer

encoding schemes greatly influence the learning capacity of transformer models and can be optimized for different genomic tasks. They conducted extensive benchmarking across multiple datasets, demonstrating that alternative k-mer schemes yield notable improvements in model generalizability and accuracy. The research contributes to a deeper understanding of pre-training dynamics for biological sequence modeling.

Simmel et al. [57] reviewed the broader implications of nucleic acid strand displacement reactions, connecting DNA nanotechnology to translational regulation. Though not a machine learning study per se, it contextualized how such molecular interactions could benefit from predictive models like DNABERT. The review highlighted the integration of biochemical and computational methods to model dynamic regulatory mechanisms. This foundational understanding serves as a springboard for applying AI models in synthetic biology and gene expression control.

Elsheikh et al. [58] proposed BERT-DNA, a transfer learning framework for identifying functional genomic regions. The model utilized transformer networks pre-trained on DNA sequences, enabling it to detect promoters, enhancers, and other regulatory elements with high precision. Their results showed substantial improvements in recognition accuracy compared to standard CNN and RNN models. The work emphasized the strength of domain-specific pre-training and the reusability of learned features across various biological tasks.

Luo et al. [59] developed iEnhancer-BERT, a model that combines DNA language modeling and transfer learning to classify enhancers and predict their strengths. By leveraging pre-trained DNABERT embeddings and fine-tuning them on enhancer datasets, the model achieved robust classification performance. Their framework underscored the potential of transfer learning in genomics and

demonstrated effectiveness in understanding the functional relevance of non-coding DNA regions.

Leksono et al. [60] applied DNABERT for splice site prediction in Homo sapiens DNA, highlighting its advantages over traditional sequence labeling models. They explored how pre-trained contextual embeddings from DNABERT could enhance recognition of canonical and non-canonical splice sites. Their evaluation across multiple benchmarks showed improved precision and recall, suggesting DNABERT's potential in transcriptome annotation and RNA splicing research.

2.4 REVIEW ON GENOMIC APPLICATIONS

Babukhian et al. [1] conducted a comprehensive analysis of the regulatory potential hidden within 5' untranslated regions (5'UTRs) by applying DNABERT-2, a transformer-based model pre-trained on genomic sequences. Their study sought to uncover cryptic regulatory elements often missed by conventional computational approaches and benchmarked DNABERT-2 against convolutional neural networks (CNNs), which are commonly used in sequence analysis. The authors demonstrated that DNABERT-2 excels in capturing long-range dependencies and subtle sequence features that CNNs typically overlook due to their local receptive fields. Through interpretability techniques like attention visualization, they highlighted how transformer models could uncover dispersed and non-canonical regulatory signals across 5'UTRs. This work is significant because it bridges the gap between machine learning performance and biological insight, particularly in regions traditionally considered difficult to annotate due to their high variability and complex regulatory architecture.

Li et al. [4] developed misORFPred, a machine learning model focused on mining translatable small open reading frames (sORFs) from

plant pri-miRNAs. By employing a scalable k-mer encoding technique along with a dynamic ensemble voting strategy, their approach significantly enhances the prediction of translatable sORFs, which are notoriously difficult to detect due to their short length and non-canonical structures. This method outperforms traditional sequence-alignment techniques and deep learning models by integrating structural and sequence-based features. The study holds great promise for understanding non-coding RNA regions and the hidden proteomic potential within plant genomes.

He et al. [10] proposed a hybrid deep learning model combining DNABERT-2 with a Bidirectional Gated Recurrent Unit (BiGRU) network for enhancer prediction. By leveraging multi-species genomic data, the model improves enhancer identification accuracy and generalizability across different organisms. Their results suggest that combining transformer-based contextual embedding with sequence modeling techniques like BiGRU provides a powerful framework for non-coding regulatory element prediction.

Li et al. [11] developed msBERT-Promoter, a two-stage deep learning predictor for identifying DNA promoters and classifying their strengths. Built on a BERT pre-trained model, it utilizes a multi-scale ensemble strategy to capture promoter features at varying resolutions. The system significantly outperforms traditional CNN and RNN-based methods and provides an interpretable scoring system for promoter strength. The work holds implications for gene expression modeling and synthetic biology applications.

Zhao et al. [19] applied BERT-based approaches to identify long non-coding RNAs (lncRNAs), which are vital in gene regulation and disease progression. Their model demonstrated superior classification performance and was capable of discerning biologically relevant sequence signals. This study illustrates the strength of contextual embeddings for analyzing poorly annotated genomic regions.

Huang et al. [43] developed pangenome-informed language models aimed at privacy-preserving synthetic genome sequence generation. Their models accounted for genomic diversity while ensuring data anonymity, a growing concern in genomic data sharing. Using DNABERT as the foundation, they incorporated constraints based on pangenomic variations to maintain biological realism. The synthetic sequences retained statistical properties of real data while preventing re-identification risks. This innovation is pivotal for developing secure genomic databases, especially in clinical genomics. Their work opens a pathway to ethically share data while preserving scientific utility. It marks an intersection of AI, privacy, and public health genomics.

Li et al. [44] conducted a study predicting phosphorylation sites related to SARS-CoV-2 infection, analyzing how these modifications may intersect with lung cancer pathways. Their DNABERT-based pipeline identified site-specific motifs in viral and host genomes. The findings suggest a regulatory overlap that could explain disease progression in co-morbid conditions. By integrating DNABERT's sequence embeddings with biological annotation databases, the model offered insights into phosphorylation-mediated signaling. The results may aid in targeted therapies or diagnostics for viral-oncology interactions. This study bridges virology and oncology using language model insights.

Zhou et al. [47] introduced DNABERT-2, a refined model capable of efficiently handling multi-species genome data. Improvements included architectural optimizations for scalability and a benchmark suite to test generalization across species. DNABERT-2 showed superior accuracy in regulatory element prediction compared to its predecessor. It highlighted evolutionary conservation in genomic features, making it a versatile tool for comparative genomics. Their model addressed the need for cross-species annotation tools in

evolutionary biology and agricultural genomics. DNABERT-2 represents a foundational shift toward more adaptable genomic models.

Tan et al. [61] conducted a comprehensive evaluation of DNABERT during the fine-tuning phase for predicting DNA sequence binding specificities. Their results indicated that fine-tuning pre-trained models on specific genomic tasks yields significant accuracy improvements over training from scratch. They also compared different fine-tuning strategies, identifying best practices for model adaptation to various prediction problems in genomics.

Viljamaa et al. [62] investigated transformer networks in gene prediction, showcasing their ability to outperform legacy HMM-based gene finders. They demonstrated how attention mechanisms can effectively model long-range dependencies within genes, leading to better exon-intron boundary detection. Their work contributes to the refinement of gene structure annotation using deep learning.

Zhang et al. [63] applied fine-tuned BERT models to predict protein-DNA binding sites, focusing on understanding molecular interactions through contextual sequence modeling. Their model achieved high performance metrics, especially in identifying binding motifs and regions critical for transcriptional regulation. The approach demonstrated the value of pre-trained models in structural genomics and molecular biology.

Palés Huix et al. [64] investigated knowledge distillation methods to compress DNABERT for efficient genomic element prediction. By transferring knowledge from large models to smaller, more efficient networks, they retained high predictive performance while reducing computational costs. Their work is significant for real-world deployment of DNABERT-based tools in genomics.

Le et al. [65] introduced BERT-Promoter, an improved sequence-based predictor of DNA promoters using BERT with SHAP-based feature

selection. This model combined interpretability with predictive accuracy, offering insights into the contribution of each nucleotide feature. Their results highlighted the synergy between explainable AI and deep learning for biological sequence analysis.

Zheng et al. [66] introduced BERT-Genome, a pre-trained BERT model specifically designed for genome sequence classification. Their approach utilized the deep learning capabilities of transformers to analyze vast genomic datasets, demonstrating significant improvements in classifying genomic sequences over traditional machine learning models. By pre-training the BERT model on genomic sequences, the authors leveraged the power of contextual embeddings to better understand DNA sequences' intricate features. This work contributed to the growing trend of applying natural language processing (NLP) techniques in genomics, paving the way for more efficient sequence classification methods in bioinformatics.

Rahman et al. [67] explored fine-tuning BERT for predicting mutations in non-coding regions of DNA, an area that is critical for understanding genetic diseases and disorders. They emphasized the importance of contextual information in non-coding regions, where mutations can have significant impacts on gene regulation and expression. Their work demonstrated that BERT's pre-trained embeddings could be effectively adapted to identify potential mutations in these regions, showcasing the model's ability to enhance mutation detection without needing extensive manual feature engineering. The study highlighted the potential of transformer models in precision medicine applications, especially in predicting mutations that affect gene regulation.

Qiu et al. [68] applied BERT-based deep representation learning to genomic sequence classification, showing the model’s power in capturing the complex patterns inherent in biological sequences. Their research focused on leveraging deep learning to process large-scale

genomic data, improving the accuracy of various sequence classification tasks. By using BERT's pre-trained weights, the model was able to handle the variability and complexity of genomic sequences, which is typically a challenge in genomics. The results underscored the increasing role of transformer models in bioinformatics and their potential to surpass traditional methods in sequence-based prediction tasks.

Nakamura et al. [69] developed transformer-based neural models for classifying disease-associated mutations in human DNA, providing a powerful tool for understanding the genetic underpinnings of diseases. They focused on using transformer networks to capture both short- and long-range dependencies in genomic sequences, which are crucial for identifying mutations that lead to diseases. Their model demonstrated improved performance in identifying pathogenic mutations compared to conventional models, making it an important advancement in the field of medical genetics and personalized medicine.

Venkatesan et al. [70] enhanced mutation detection in DNA sequences by utilizing multi-head attention mechanisms in transformer models. Their study highlighted the importance of attention mechanisms in analyzing genomic data, allowing the model to focus on the most informative parts of the sequence. By improving the detection of mutations, particularly those that are subtle or rare, their work has implications for genetic diagnostics, potentially helping to identify mutations linked to a variety of diseases more accurately.

Dao et al. [71] introduced FlashAttention, a fast and memory-efficient attention mechanism that significantly enhances the computational efficiency of transformer models. This work addressed the common challenge of scaling attention mechanisms to large genomic datasets, improving both speed and memory usage without sacrificing accuracy. FlashAttention's innovative approach is particularly useful in large-scale genomic studies, where the volume of

data can be a bottleneck in processing. This research contributes to making transformer models more practical for real-world genomic applications by making them faster and more efficient.

Stachowicz et al. [72] explored the combination of DeepSEA CNN and DNABERT for regulatory feature prediction in non-coding DNA regions, where most genetic regulation occurs. His study focused on using these models to predict functional genomic elements, such as enhancers and promoters, in non-coding regions of DNA, which play critical roles in gene expression. The results demonstrated that combining CNN and transformer models like DNABERT could lead to more accurate predictions, providing insights into the complex regulatory mechanisms of the genome.

Le et al. [73] proposed a hybrid architecture combining BERT and a 2D convolutional neural network (CNN) for identifying DNA enhancers from sequence information. Their model utilized the strengths of both transformer-based and CNN architectures to process genomic sequences in a way that captured both global sequence features and local positional dependencies. This approach proved particularly effective in enhancer prediction tasks, where the ability to capture long-range sequence dependencies and local sequence motifs is crucial for identifying regulatory elements accurately.

Bressem et al. [74] utilized deep learning natural language models pre-trained on a large corpus of chest radiographic reports to classify medical images. While not directly related to genomics, their work introduced innovative methods for applying deep learning to domain-specific text, which has parallels in genomic data analysis. Their study demonstrated the effectiveness of large-scale pre-training for classification tasks, providing insights that could be transferred to genomic sequence analysis and other bioinformatics applications where labeled data is limited.

Tang et al. [75] focused on DNA strand displacement reactions as a method for discriminating single nucleotide variants (SNVs). Although this work is more focused on the biochemical side, it presented a valuable tool for distinguishing between variants at a very fine scale. Their approach has implications for DNA sequencing technologies and diagnostic applications, showing how biochemical techniques can complement computational methods, like those used in DNABERT, to enhance the accuracy of variant detection. This combination of experimental and computational techniques is crucial for advancing genomic research.

2.5 RESEARCH GAPS

Although deep learning techniques have partially solved problems faced by DNA sequence analysis, there are still various factors impeding the accuracy and applicability of mutation classification models, and this is particularly true in the analysis of the viral genome. The following research gaps have been identified that underline the need for improving classification accuracy, biological interpretability, and cross-domain transfer:

- **Generalization across Viral Species:** Current models are frequently unable to generalize mutation classifications across different viral species because of differences in genomic structures and sequence patterns. Whereas DNABERT-2 allows for some support across multiple species, it does not reliably generalize its learned representation to rare or obscure viral genomes. There is a need for more robust pretraining strategies and adaptive fine-tuning methods.
- **Insufficient Fine-Grained Mutation Classification:** Many studies focus on broad mutations or binary classifications rather than fine distinctions between different mutations (e.g., distinguishing synonyms from non-synonyms or insertions from deletions). This disregard for finer mutation distinctions

minimizes the biological relevance of their predictions and restricts application areas like vaccine development or resistance analysis from these mutations.

- **Limited Integration of Temporal Context and Sequential Dynamics:** While traditional transformer models like DNABERT-2 successfully capture local and contextual features, they often fail to capture the sequential dynamics of mutation evolution. LSTM models can do this, the use of LSTMs together with transformer-based representations has not yet been extensively explored for genomic tasks, paving the way for the further development of hybrid models that exploit global attention and temporal alignment.

  True, Continued efforts on developing DNA sequence analysis using deep learning techniques have much assisted the mutation classification model, though many challenges still affect the accuracy and applicability of the model in the analysis of viral genomes. Research gaps among which the need to enhance classification accuracy as well as improve biological interpretability and ability to transfer across domains interpret the following points:

- **Generalization across Viral Species:** Generalization of mutation classifications across viral species still remains challenging in interpretations by the current models mostly due to genomic structural differences and sequencing patterns. Although DNABERT-2 supports multiple species, it still suffers when trying to reuse learned representations for their adaptation to rare or less common viral genomes. Hence, stronger pretraining strategies and adaptive fine-tuning methods will be imperative.
- **Low Resolution Fine-Grained Mutation Classification:** Most studies accomplish mutation detection or binary classifications but do not account for any kind of granularity in the category (e.g., synonymous vs. non-synonymous mutation or insertions vs. deletions). Thereby, diminishing their biological importance of

the prediction and creating restrictions in addresses, for example, vaccine development or analysis of drug resistance.

- **Minimal Integration to Temporal and Sequential Context:** Traditional transformer models such as DNABERT-2 may perform well in capturing local features as well as contextual features. Often time evolution of mutations remains considerable lagging behind. Hybrid models with both global attention and temporal dependencies are needed since it hasn't been much researched on coupled LSTM with transformer-based representation.
- **Lack of Domain-Specific Annotations:** Most viral genome datasets lack good-quality labeled mutations, and this has made the construction of supervised learning problems hard. Existing models often rely on large repositories of general genomic datasets which may not capture task-relevant mutation signals. There is a need for weakly supervised or semi-supervised methods that can perform well with limited noisy, or domain-specific labels.
- **Evaluation on Real-World Viral Datasets:** Present-day mutation classification models are generally tested on curated or synthetic datasets that do not accurately reflect the real-world diversity and complexity of genomes. For the established performance in real-life scenarios, the models need to be evaluated with noise-imbalance, and heterogeneous sources of viral sequences from databases like NCBI Viruses.
- **Model Interpretability and Biological Validation:** Although some high-accuracy predictions can be made by deep learning algorithms, their predictions often lack biological interpretability. Connecting what is learned to the known biological pathways, mutation hotspots, or phenotypic outcomes is crucial for building trust and applicability in clinical or epidemiological settings.

To address all these mentioned concerns, the project proposes a new hybrid model based on the combination of DNABERT-2 and different LSTM architectures, interspersed with fine-tuned k-mer embeddings, mutation-aware attention mechanism, and thorough evaluation on multiple viral datasets. The general aim is, thus, to improve mutation classification performance but with real biological meaning and applicability across domains.

2.6 SUMMARY

The overview of literature highlights some of the transformations and trends associated with the classification of mutations in viral genomic sequences as they touched on advancements made and challenges that keep remaining in mutation classification. Basic initial approaches based on traditional machine learning had serious limits in comprehending the highly complex and high-dimensional nature of nucleotide sequences and lacked robustness across different viral genomes. The emergence of effective deep learning models, such as DNABERT-2, which based on a transformer framework, has greatly improved sequence representation in terms of k-mer embeddings used to assist in mutation detection and classification. These contextually strong models still demonstrate difficulties in sequential dependency modeling and generalization across species.

A lot more big issues are currently being faced in ongoing research, including insufficient hybrid models that integrate both the contextual and temporal aspects, high dependability on large annotated datasets for training, and limited interpretability pertinent to real-world biological applications. Most models fail to adequately represent dynamic patterns of mutation, specifically when influenced by small, high inter-class similarity or noise datasets. This study aims to fill such gaps through DNABERT-2 for rich contextual DNA feature capture and LSTM networks to learn sequential changes in nucleotides with time.

CHAPTER 3

METHODOLOGY

This chapter details the research design, methods, and procedures used to conduct the study. It explains the data collection process, tools and technologies employed, and the rationale behind selecting specific techniques. The methodology ensures reproducibility and clarity, outlining how the objectives of the study are systematically approached. It also includes information on experimental setups, algorithms, or frameworks used during implementation.

3.1 INTRODUCTION

The latest initiatives in the domain of deep learning have recorded significant success in genomic sequence analysis. Nevertheless, distinguishing mutations across different viral genomes is a puzzle owing to the intricacies involved in handling high-dimensional data, sequencing dependencies, and variabilities. Machine learning methods operate mostly in a low-dimensional regime and fail to maintain long-range dependencies, while the better current models may have troubles generalizing to different viral species or adapting to subtle mutation patterns. To further this goal, this work proposes a hybrid architecture that combines DNABERT-2, a transformer-based model for genomic sequences, and LSTM networks.

DNABERT-2 exploits k-mer tokenization to identify local and global contextual patterns in nucleotide sequences, whereas LSTM enhances temporal modeling through learning dependencies and transition of mutations in the sequence. Such a combined approach will enhance classification accuracy, generalizability, and biological interpretability in identifying viral mutations. Combining the strengths of both transformer and recurrent architectures, the proposed model offers a robust solution to tackle the challenges presented by real-world

genomic data, especially in the area of cross-species mutation classification over datasets such as NCBI Viruses.

3.2 DATASETS UTILIZED

The datasets of nucleotide from NCBI Virus compile all viral sequence data from GenBank and other sources. These datasets can directly be helpful for natural language processing (NLP) applications in genomics-most importantly for classifying mutations through transfer learning using languages such as DNA-BERT and LSTM.

- **Human Viruses Dataset-22,463 Entries**
  The dataset contains nucleotide sequences of human infecting viruses. In genomic studies, this data can help scientists study viral mutations, evolutionary trends, and host/virus interaction. The researchers can apply the DNA-BERT model, which is basically a transformer model, to the genomic data as a valuable genomic insight tool for mutation classification and disease prediction. At the same time, the LSTM (Long Short-Term Memory) networks enable sequence processing by modeling long-term dependencies within the viral genome. This dataset is helpful in mutation tracking, which plays an essential role in vaccine development and developing antiviral drugs.

- **Influenza Viruses Dataset-14,041 entries**
  This dataset represents nucleotide sequences among different strains of Influenza A and B viruses. Influenza viruses are responsible for seasonal outbreaks and pandemics due to their ability to mutate rapidly. DNA-BERT models in NLP sequencing enhance the classification of mutations based on recognizing, while the temporal dependencies examined by LSTM networks improve predictions in the viral evolution. The

dataset is crucial in monitoring influenza and allowing researchers to predict new strains with which they can hopefully develop vaccines. Genetic analysis has been significantly advanced through transfer learning-based models, providing a deeper comprehension of antigenic drift and shift.

- **SARS-CoV-2 Dataset of Viruses (3,506 entries)**
  The SARS-CoV-2 dataset lies within the nucleotide sequences of the virus that caused COVID-19. Given that this virus has a specific rate of mutation so much so that one would find it in an entirely new variant by the time he or she gets to the final stages of his or her studies, this dataset is of great value in tracking variants and evaluating their transmissibility and effects on immune evasion. In mutation analysis within advanced deep learning upon a genome sequence analysis, DNA-BERT does the work, while LSTM networks capture sequence dependency in the viral evolution. This library has been instrumental in variant detection and informing public health and vaccination strategies. The application of language transfer learning models in genomics has totally changed SARS-CoV-2 research by making real-time mutation tracking and predictive modeling a practical reality.

The integration of these datasets hence pushes genomics research forward, providing foundations for mutation classification, vaccine development, and epidemiological studies. The combination of DNA-BERT with LSTM should unravel viral evolution in order to formulate better strategies for preventing and treating diseases.

3.3 LIBRARIES UTILIZED

Pandas is a robust library in Python for data manipulation and analysis. This particular library is widely used for handling large

nucleotide data sets, especially in genomics. In case of NLP sequencing, pre-processing genomic sequences becomes easier by converting raw data into DataFrames through Pandas, allowing users to filter, clean, and modify nucleotide sequences for mutation classification. It makes the processing and analysis of mutations easier with language transfer learning models by pairing it with DNA-BERT, organizing genomic text data into tokenized formats. In addition, Pandas is used to do batch processing, which enables a researcher to work with extensive genomic datasets efficiently. This is complemented by its compatibility with LSTM networks to analyze sequential data while leaving long-range dependencies () to be maintained in DNA sequences.

NumPy is one of the most basic libraries done for numerical computing with Python. A vast amount of genomic work has happened using this library as it supports efficient matrix operations and a wide variety of array manipulations. From the perspective of natural language processing (NLP) sequencing, NumPy is helpful in converting nucleotide sequences to their corresponding numerical formats very crucial for mutation classification using DNA-BERT and LSTM. Managing multi-dimensional arrays makes it a good fit for encoding genomic sequences into appropriate embeddings for deep learning models. Lastly, NumPy makes computations faster so that language transfer learning models analyze large-scale genomic datasets efficiently. Its compatibility with TensorFlow and PyTorch improves model training, allowing researchers to create a strong base towards mutation classification.

The **Transformers** library made by Hugging Face is at the heart of language transfer learning models for genomics. It has pretrained models, such as DNA-BERT, primarily aimed at working with nucleotide sequences. For NLP sequencing tasks, researchers can now utilize self-attention mechanisms embedded in Transformers to understand more sophisticated relationships in genomic data, improving mutation classification accuracy. The library also provides fine-tuning

functionality, which allows scientists to adapt pretrained models to certain genomic datasets. The transformers also assist with sequence embedding, which guarantees that DNA sequences are well described for deep learning models, such as LSTM, in analyzing sequences. Both PyTorch and TensorFlow can be supported, making it a very flexible resource in genomics.

With the help of **PyTorch**, you will be able to develop, train and optimize the models. You can run a handful of tons of genomic data efficiently because the torch library will serve as a backbone for operations of tensors and GPU acceleration. Because it has the important Dataset and DataLoader classes, torch.utils.data plays an important role in managing datasets, as it provides ability for batching, shuffling, and parallel loading of data that are relevant when training over large genetic sequences. The torch.nn module provides the components required to create deep learning models, such as built-in layers nn.Embedding that can be used to represent sequences, nn.LSTM which enables the model to remember long-range dependencies in DNA sequences, and nn.Linear for purposes of classification. DNA-BERT is a transformer-based architecture which processes k-mer tokenized sequences and extracts prominent features for capturing sequential dependencies through an LSTM layer before this is classified. The torch.optim module focuses more on model optimization via algorithms like Adam and SGD towards the minimization of loss functions in terms of weight adjustment through backpropagation. Optimization for a genomic model is important due to the massive scale of the sequence embeddings that are often handled and mutation patterns that become very complex.

This model combines bidirectional DNA-BERT's contextual embeddings to assist with the LSTM-based classifier, which effectively allows sequential information to be conveyed through recurrent connections. The pipeline includes torch.utils.data loading genomic sequences, using DNA-BERT that obtains feature extraction from those

sequences, feeding them into the LSTM to catch long dependencies, and finally applying a fully connected layer to classify mutations. The process of optimization is now in the hands of torch.optim, whose function creates the appropriate improvements in model parameters that will lead to a better outcome. Thus, improving its classification while making it generalizes quite well across different viral mutations. The combination of contextual knowledge from DNA-BERT and sequential learning from LSTM increased the detection of complex mutation patterns, thus achieving very high accuracy with robustness in genomic NLP applications.

3.4 AUTOTOKENIZER

For the mutation classification task in genomics strengthened by NLP methodologies, models such as DNA-BERT and LSTM implemented with AutoTokenizer pre-process DNA sequences from the Hugging Face transformers library. DNA sequences are extremely encoded with nucleotide basis: A, T, G, or C, and they require tokenization other than natural languages to make them optimal for deep learning modeling. Raw nucleotide sequences into tokenized formats defined according to the model's input requirements would be done seamlessly through it with a pre-trained DNA-BERT model. A DNA-BERT would use k-mer tokenization instead of generating separated character tokens; it breaks the DNA sequences into overlapping k-length substrings. This kind of genetic sequences that consist of huge genomic datasets like NCBI viruses does not require manual preprocessing but directs the automated process of maintaining consistency. More so, this is how easy and effective DNA sequences can automatically be numerically represented for embedding layers of the DNA-BERT.

These sequences are converted after tokenization into input IDs and attention masks for appropriate processing of the transformers over data. The model's vocabulary comprises these numerical indices referred to as input IDs corresponding to the tokens, while attention

masks refer to actual tokens vis-a-vis padding of the model so that it does not perform unnecessary calculations concerning the values. Such a scenario might arise when training models on genomic sequences of various lengths. The resulting padding will ensure that batch sizes are similar, while the attention mechanism ensures that models do not learn the padding tokens invented artificially.

The AutoTokenizer facility provides seamless integration with DNABERT that allows users to fine-tune models over custom genomic datasets without having to go through the complexity of manually transpires sequences. When processed alongside an LSTM classifier, the tokenized sequences of mutation patterns from DNABERT's embeddings are further sent through an LSTM layer where the processor fetches the long-range dependencies in the mutation pattern. AutoTokenizer efficiency of well-structured tokenized sequences guarantees high-end input representation offered to the transformer-based DNABERT model leading to improvement in classification performance. Considering that AutoTokenizer would still support raw BERT versions, its shared adoption promotes transfer learning as new research can narrow down to fine-tuning models on relevant genomic tasks beyond mutation classification. It also supports batch tokenization, which drastically increases the speeds at which data can be preprocessed to handle millions of sequences-very important for high-throughput genomic research. Further, special tokens generated through AutoTokenizer, [CLS] for classification tasks and [SEP] for separating sequences, provide the model with better context because it can understand the relevant aspects of the sequence as a whole. Genome researchers can thus completely automate the entire NLP workflow with input and output of information sequences-ensuring reproducibility, scalability, and high efficiency in genomic mutation classification tasks.

3.5 POSITIONAL ENCODING

The positional encode is typically represented as a matrix which is summed up with the embeddings used in the transformer architecture. Creating positional options is possible in various ways, out of which the one most-acquainted employs sine and cosine functions to calculate their values.

1. Positional Encoding Matrix: The positional encoding matrix, PE, is created for an input sequence of length L and embedding dimension d. Thus the positional encoding matrix will be of size L×d that is L rows and d columns. Each row of the matrix is supposed to encode the positional encoding vector for the corresponding token in the sequence.
2. A positional encoding with sine and cosine functions calculates the positional encoding vectors for each token.

$$\boldsymbol{PE_{pos,2i}} = \mathbf{sin}\left(\frac{\boldsymbol{pos}}{\mathbf{10000}^{\frac{\mathbf{2i}}{\boldsymbol{d}}}}\right) \tag{3.1}$$

$$\boldsymbol{PE_{pos,2i+1}} = \mathbf{cos}\left(\frac{\boldsymbol{pos}}{\mathbf{10000}^{\frac{\mathbf{2i}}{\boldsymbol{d}}}}\right) \tag{3.2}$$

In this context, "pos" implies the position of the token in the sequence, "i" is the index within the embedding dimension, while "d" represents the size of the embedding space.

The PE matrix is applied to the input embeddings of the transformer model through element-wise operations before being supplied to the network for further processing. It is an important step as it offers crucial positional information to the model about each token in the sequence.

The importance of positional encoding, in this case, is vital to the ability of the model to comprehend and process sequential data. It provides little-known insights into how a token is placed from other tokens in the input sequence, thus giving the model the sequential context and dependencies for various tasks: translation, text

summarization, or sentiment analysis. Positional encoding is one of the keys to how transformers work, a specific type of model for various natural language processing tasks. By conveying critical information about the relative positioning of tokens in input sequences, it situates itself amid the transformers' self-attention capability and the actual sequential order of the data. Thus, positional encoding will remain a core element of all transformer arrangements as the field of NLP continues to advance toward deeper insights and generation of natural language.

3.6 SINGLE-HEAD ATTENTION MECHANISM

In a single-head attention model, multiple attention mechanisms work in parallel to capture various types of patterns. It uses a single set of query, key, and value matrices to compute attention scores. In this set-up, input tokens (words or subwords) are first mapped to continuous vectors, and these are then converted into queries, keys and values via learned weight matrices. The attention scores are calculated between the dot products of the queries and the keys and then normalized with a Softmax function to reflect how much relative importance each word has in the sequence.

Dot products of the value vectors are taken, which constructs a context vector with each token containing the information relevant to that token. Thus, while single-head attention is less expensive, it is not multi-headed. Single-head attention can still exceed expectations where smaller, more resource-constrained models are concerned: simple and efficient will work better for translation in that specific case of transformer architecture. And now let us look into how a single-head self-attention step looks like:

Input Embeddings: An input sequence is $X = \{x_1, x_2, x_3, \ldots, x_n\}$.

Query, Key, and Value Vectors: Using these three vectors associated with each input element x_i , each element will have a query vector q_i, a key vector k_i, and a value vector v_i. These vectors are derived from input embeddings and inputs to the attention-aware score generator.

Attention Scores: The attention score between an input pair of x_i and x_j is the score generated through dotting the corresponding query and key vectors. d represents the dimensionality of the vectors involved in the aforementioned operations.

$$a_{ii} = softmax(q_i . \frac{k_i}{\sqrt{d}}) \tag{3.3}$$

Weighted Sum: Once having the attention scores, they are used to compute a weighted sum of the value vectors:

$$Attention(X) = \sum_{j=1}^{n} a_{ij} . v_j \tag{3.4}$$

This weighted sum reflects the result of the single-head self-attention process for the input sequence X.

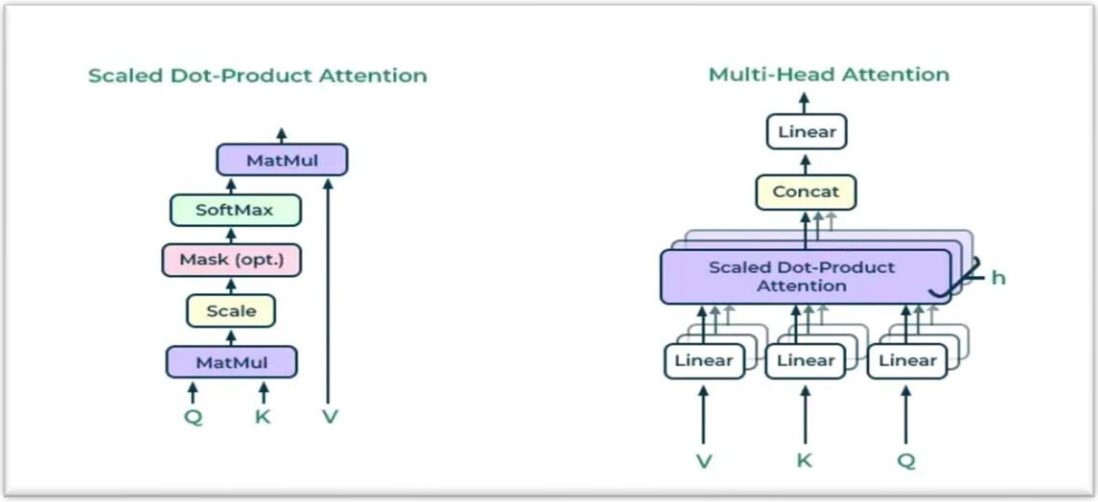

Figure 3.1 (a) Single-Head Mechanism (b) Multi-Head Mechanism

The Figure 3.1 represents the architectures of attentions mechanisms to understand input embeddings.

The attention mechanism, originally introduced in the context of machine translation, has significantly transformed the way deep learning models process sequential data, including DNA sequences. In the domain of computational biology, DNA sequence processing involves analyzing nucleotide strings composed of adenine (A), thymine (T), cytosine (C), and guanine (G). These sequences encode the genetic blueprint of living organisms and are fundamental to understanding genetic regulation, mutation, expression, and other biological processes.

Traditional models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) often faced challenges in handling long-range dependencies in sequences. For instance, biological functions such as enhancer-promoter interactions or transcription factor binding can involve interactions between nucleotides separated by hundreds or thousands of bases. These long-range dependencies are often not well captured by traditional fixed-size filters (CNNs) or by the vanishing gradient limitations of RNNs. The attention mechanism addresses this issue by enabling the model to weigh and focus on different parts of the input sequence dynamically.

Instead of processing input in a strictly sequential or fixed-window manner, attention allows the model to assign relevance scores to each position in the sequence relative to a query, facilitating a global view of the sequence. This is crucial in DNA sequence analysis because important biological signals or motifs can appear in various positions and may not follow strict positional rules. The attention mechanism thus offers a flexible, data-driven way to learn which regions of the DNA sequence are most influential for a specific task, such as classification of mutation types, identification of regulatory elements, or prediction of gene expression levels.

The function of attention mechanisms in DNA sequence processing involves computing a weighted representation of the entire

sequence based on the importance of each nucleotide or k-mer (a substring of length k). In a transformer-based model like DNABERT, sequences are first tokenized into overlapping k-mers (e.g., 6-mers), and then each token is embedded into a high-dimensional vector space. Once embedded, the attention mechanism comes into play by evaluating the relationships between all pairs of tokens using a combination of queries (Q), keys (K), and values (V).

For each token, a query vector is compared to key vectors from all other tokens to produce an attention score, which represents the relevance of one token to another. These scores are then normalized using a softmax function and applied to the corresponding value vectors to produce a context-aware representation of the token. This process allows the model to attend to biologically significant regions of the DNA, regardless of their position in the sequence. For example, in identifying a mutation that leads to disease, the attention mechanism might highlight not only the mutation site itself but also upstream or downstream regulatory regions that interact with the mutation site.

This is especially helpful in capturing motifs like transcription factor binding sites or splice sites that influence gene function. Moreover, attention mechanisms can operate in multiple heads, allowing the model to learn different types of dependencies or interactions in parallel, which is essential given the complexity of biological systems. By using self-attention, the model essentially builds a dynamic interaction map of the sequence, where each part of the DNA can potentially influence and be influenced by every other part, providing a comprehensive understanding that is particularly suited for tasks like sequence classification, promoter prediction, and motif discovery.

The integration of attention mechanisms into DNA sequence models has led to significant advancements in genomics and bioinformatics research. One of the most notable advantages is the

interpretability that attention provides. In contrast to traditional deep learning models, where it is often difficult to trace why a certain decision was made, attention weights can be visualized to highlight which parts of the sequence the model focused on during prediction. This not only aids in model validation but also provides biological insights by potentially uncovering novel functional elements or interactions within the genome. For instance, when used in enhancer–promoter interaction prediction, attention can reveal previously unknown distal regulatory elements that influence gene expression.

Furthermore, attention mechanisms enable models to generalize better across different datasets and organisms, as they allow for adaptive focusing on relevant sequence regions rather than relying on fixed positional biases. This generalization is crucial for tasks such as cross-species prediction of gene regulatory elements or the identification of conserved motifs. Additionally, the scalability of attention-based models makes them suitable for large genomic datasets, which is often a limitation for traditional models due to computational constraints.

The use of pre-trained transformer models like DNABERT leverages unsupervised learning on massive DNA corpora, enabling fine-tuning on smaller datasets for specific tasks, thereby reducing the need for extensive labeled data. This transfer learning capability, combined with the nuanced focus provided by attention, significantly enhances performance in tasks such as variant effect prediction, DNA–protein binding affinity estimation, and epigenetic feature detection. In summary, attention mechanisms empower DNA sequence models with a powerful blend of contextual understanding, long-range dependency capture, and interpretability, ultimately advancing the field of computational genomics by enabling more accurate, robust, and biologically meaningful predictions.

3.7 MULTI-HEAD ATTENTION MECHANISM

The multi-head attention mechanism is an important feature in the models of a transformer for enabling a language into effective translations because it brings forth different aspects of relationships between words in a sentence. Unlike single-head attention, which may look only for one type of linguistic feature, multi-head attention can have several attention heads to identify different patterns, for example, syntactic dependencies and semantic similarities, at the same time. While each attention head computes a scaled dot-product attention independently by transforming the input embeddings into query, key and value vectors using learned weight matrices, they can work well alone in concentrating on areas of the input sequence to afford better understanding of the word contexts in parallel. As such, outputs from all heads will be concatenated and linearly transformed to serve this purpose. It is an encoder-decoder architecture that will have the encoder focusing on the relationships within source text and the decoder will predict the target text attending to relevant outputs from the encoder. This is how it allows smaller models to construct better fluency and accuracy translations under limited resources.

Input embeddings: like the single-headed variant starts with an input sequence $X=\{x_1,x_2,\ldots,x_n\}$, where each x_i represents an element in the sequence.

Query, Key, and Value Vectors: For each input element x_i, different sets of query, key, and value vectors are generated for one attention head. These vectors will be the derived vectors from the input embeddings and hence capture different dimensions of the input sequence.

Attention Scores: For each attention head h, attention scores α_{ij}^{h} are computed for pairs c and x_j based on their respective query and key vectors as,

$$a_{ij}^{h} = softmax(q_i^h . \frac{k_j^h}{\sqrt{d_h}}) \quad (3.5)$$

Here, d_h represents the dimensionality of the query and key vectors for the h-th attention head.

Weighted Sum: Computes a weighted sum of the value vectors using the attention scores for each attention head:

$$Attention^{h}(X) = \sum_{j=1}^{n} a_{ij} . k_j \quad (3.6)$$

This results in several output vector sets with each set linked to a distinct attention head.

Concatenation and Projection: In the final phase, all the output vectors from all the attention heads are combined and projected into the original output space.

3.8 BERT ARCHITECTURE

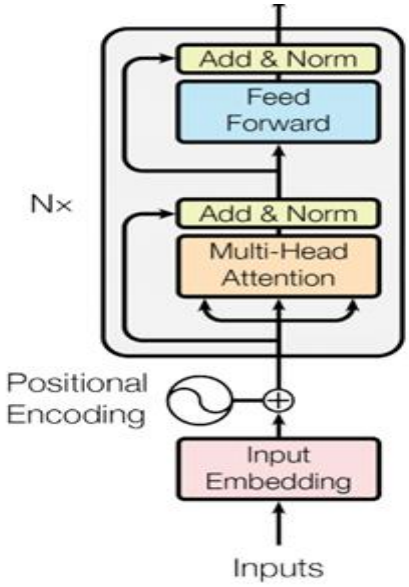

Figure 3.2 BERT Architecture

The above Figure 3.2 describes the flow of encoding in BERT process.

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a new deep learning model from Google that revolutionizes natural language processing (NLP) through the advent of bidirectional context awareness for end-user applications. Unlike conventional models that read text in a single direction, BERT analyzes text from both ends - backward and forward - such that it understands the context of a word in relation to all the other words in a sentence. It brings such a significant improvement in almost everything, from text classification and question-answering to named entity recognition, because it captures complex relationships between words that a unidirectional model would not be able to discover.

BERT consists of stacks of tranformers with each blocks lincating multi-head self-attention and feed-forward networks. The model achieves this by accommodating multiple sentences on simultaneous attention of segmentations, learning very different linguistics as well as dependencies between words. Feed-forward layers then improve the interpretation of a word but with respect to this output. This includes within each layer residual connections and normalization (Add & Norm) into stabilization and performance improvement during the learn process. In addition, BERT entails pre-trained massive text datasets with two unique training principles, namely masked language modeling (MLM) and next sentence prediction (NSP). MLM is such that some words are hidden randomly, although the model can be trained to predict them with embedded context as NSP facilitates BERT to run through relationships between a sentence pair - such that both prove vital to its strong language comprehension.

Essentially, what BERT has is the capability for fine-tuning for specific NLP tasks without requiring a long retraining. Models like DNA-BERT use the same transformer-based framework to analyze nucleotide sequences in specific fields like genomics and mutation classification, effectively identifying differences within each of such sequences. LSTM networks accompany BERT to manage long-range dependency in

genomic sequences, thus improving mutation classification work accuracy. So far, BERT has gone into some applications beyond standard NLP in advancing such domains as health, biomedical research, and genetics, by opening up opportunities for language transfer learning models. Its learning of understanding very complex problems in terms of context has made it so useful and thus established as a very important instrument of modern AI applications.

3.9 IMPACT OF DNABERT

DNABERT is a specialized adaptation of the BERT (Bidirectional Encoder Representations from Transformers) architecture, designed specifically for genomic data, particularly DNA sequences. Traditional BERT was originally developed for natural language processing (NLP) tasks, where it revolutionized contextual representation learning by using a bidirectional transformer to learn word embeddings based on surrounding context. Recognizing the similarities between human language and biological sequences, the creators of DNABERT proposed that the same principles could be applied to DNA. DNA, while not a natural language, is also a long string of characters composed of four basic nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). Just as NLP models treat words as fundamental tokens, DNABERT treats fixed-length substrings of DNA known as **k-mers** as tokens.

A k-mer is a sequence of k consecutive nucleotides (e.g., for k=6, the string "ATCGGT" is a 6-mer). DNABERT was pre-trained on massive corpora of genomic data using masked language modeling (MLM), just like traditional BERT. This involved randomly masking certain k-mers in DNA sequences and training the model to predict the masked elements based on their context. The motivation was to enable the model to learn rich contextual representations of DNA, capturing both short- and long-range dependencies across the genome. This is particularly crucial because biological functions, such as gene regulation, mutation impacts, and protein-DNA binding, often depend

on the broader sequence context—not just local motifs. The bidirectionality of DNABERT allows it to assess both upstream and downstream sequence elements when determining the functional impact of a given site, enhancing accuracy in downstream tasks such as promoter identification, splice site prediction, or mutation classification.

One of the most critical adaptations in DNABERT is the way it processes DNA sequences. Since standard BERT is built to operate on language tokens like words or subwords, DNABERT uses **k-mer tokenization** to transform nucleotide sequences into sequences of overlapping k-length substrings. For example, given a DNA sequence "ATGCGTAGC", a 6-mer tokenization would produce the following tokens: "ATGCGT", "TGCGTA", "GCGTAG", and "CGTAGC". The choice of **k** is crucial and typically ranges between 3 to 6; 6-mer has been widely adopted as it offers a balance between capturing meaningful biological motifs (which often span 6 nucleotides, like transcription factor binding sites) and manageable vocabulary size. Since there are four nucleotides, the number of unique k-mers grows exponentially as 4^k. For 6-mers, this results in 4096 unique tokens. These tokens are mapped to embedding vectors using a learnable embedding layer, just like words in NLP. A special [CLS] token is added to the beginning of each sequence to serve as a global sequence representation, and a [SEP] token can be used at the end, although its use varies depending on the task. The sequences are often padded or truncated to a fixed length to fit within the maximum sequence length supported by the model—commonly 512 tokens, as in original BERT.

DNABERT uses **positional embeddings** to encode the relative position of k-mers within the sequence, helping the model to distinguish between motifs that occur at different positions. During training, random k-mers are masked and the model learns to predict them from their context using self-attention layers. This masked language modeling (MLM) approach allows DNABERT to develop a deep

contextual understanding of genomic sequences. The model doesn't just memorize motifs but learns how their function might vary based on context—e.g., a promoter-like sequence that functions differently based on its distance from a transcription start site. These rich embeddings serve as highly informative input for fine-tuning on various biological tasks.

DNABERT adopts the **base BERT architecture** with some domain-specific modifications. The original DNABERT model is built upon **BERT-base**, which consists of 12 transformer encoder layers (also called transformer blocks), each with 12 self-attention heads and a hidden size of 768 dimensions. This results in approximately 110 million trainable parameters. The attention heads allow the model to capture multiple types of dependencies across different parts of the input sequence, which is extremely valuable in genomics where distant regions of DNA may functionally interact. Each token (i.e., k-mer) is embedded into a 768-dimensional vector, and these vectors are passed through the transformer layers, where multi-head attention dynamically weights different tokens depending on their learned relevance to the task.

The [CLS] token output from the final layer is often used as a sequence-level embedding and fed into a classification head during downstream tasks like promoter prediction or mutation classification. Fine-tuning DNABERT for a specific task involves adding a task-specific layer (e.g., a softmax classifier for multi-class prediction) on top of the pre-trained model and training it using labeled examples. During fine-tuning, the weights of the pre-trained model are updated slightly to adapt to the specific characteristics of the new dataset, while retaining the foundational genomic knowledge gained during pretraining. DNABERT also supports **multi-task learning** by allowing the same architecture to be fine-tuned for different genomic tasks, thereby promoting knowledge sharing between tasks like enhancer recognition, chromatin accessibility prediction, or methylation site detection. Key

hyperparameters for training and fine-tuning DNABERT include batch size (commonly 16–32), learning rate (typically around 1e-5 to 5e-5), maximum sequence length (often 512 k-mers), and number of epochs (commonly 3 to 10 depending on task complexity and dataset size).

The model is trained using variants of the Adam optimizer with weight decay. Pretraining is typically done on large genomic corpora like the human reference genome (GRCh38), ensuring the model learns generalized DNA patterns. By leveraging transformer-based architectures and domain-specific tokenization, DNABERT significantly enhances the capability of computational models to interpret and analyze DNA sequences, offering superior performance in various sequence-based bioinformatics tasks compared to traditional deep learning approaches.

DNA-BERT has revolutionized genomics natural language processing (NLP) by tailoring transformer models, originally endowed to human language, towards analyzing DNA sequences. Conventional models for the analysis of genetic sequences have been much dependent on manual feature engineering limited by biological assumptions. DNA-BERT resolves these challenges by using self-attention mechanisms to grasp contextual relationships in terms of nucleotide sequences and thus provide a data-driven approach to genomics analysis. By means of k-mer tokenization, DNA-BERT views DNA as a strung-together language finding salient patterns and connections in genetic sequences. This progress has furnished enormous accuracy and speed boost to a wide variety of genomic tasks, including mutation classification, promoter identification, and gene function prediction, which is evidence of its socio-political and economic importance in bioinformatics and computational biology.

Besides introducing a series of mutation classes, operational DNA-BERT contributes further to mutation classification. Identification and categorization of mutations can illuminate disease mechanisms,

viral evolution, and genetic disorders. DNA-BERT is improving performance in classification by deriving significant embeddings characterizing both local motifs and long-range dependencies that might otherwise be ignored by more classical methods. Coupled with other deep architectures, like LSTM-based classifiers, DNA-BERT adds to generalizability across many genomic datasets, particularly large viral databases like the NCBI Viruses. By assessing mutation patterns in SARS-CoV-2 variants and other viral genomes, DNA-BERT provided some pertinent evidence for epidemiological studies and vaccine development. Its ability to transfer learning into related genomic tasks also enhances understanding of genetic regulatory elements and epigenetic modifications.

Besides that, DNA-BERT laid down the foundations for language model-associated genomics and thus opened the door for large-scale transformer architectures within biology. It carries a multi-dimensional flavor by connecting NLP and genomics by extending the transformer models beyond human language to structured biological sequences. That have encouraged following fine self-supervised learning approaches on genomic data, thus reducing dependency on manually labeled datasets, and making deep learning usage in bioinformatics easier. DNA-BERT has also fueled the emergence of novel genomic transformer models, specifically designed for different sequence lengths and biological contexts, which will exaggerate the roles of NLP in genomics. DNA-BERT's have alluded to ever-becoming impacts as it begins to scale, together with an expanding data economy and data collection, with a view to more accurate, scalable, and interpretable AI discoveries in genomics for precision medicine, evolutionary biology, and synthetic biology.

3.10 PRETRAINED MODELS DESCRIPTION

Available on Hugging Face is the **Zhihan1996/DNABERT-2-117M** model, which is a deep learning pretrained model based on the transformer design and designed specifically for studying genomic sequences. It improves upon the development of previous DNABERT architecture by enhancing the model efficiency, changing pretraining methods, and improving the effectiveness of sequence representation, which has led to making it an important tool for various genomic natural language processing tasks, such as mutation classification, gene expression prediction, enhancer identification, and regulatory element detection. Contrary to deep learning models that largely depend on exhaustive manual feature engineering, DNABERT-2 depends on self-supervised learning to capture complex sequence patterns and hence generalizes across multiple genomic data sets with very little, if not negligible, labeled data. This is a moderately sized model, with 117 million parameters-a transformer that strikes a balance between computational efficiency and performance. DNABERT2 has some advances: an enhanced signature k-mer tokenization scheme that can break DNA sequences into overlapping subsequences of length k (k=3, k=6) does a better job of giving the model insight around biological motifs, mutation impacts, and long-range dependencies in genetic sequences. This tokenization scheme helps the model capture contextual relations between nucleotides, just as a BERT model would consider words as units of meaning in natural languages. DNABERT-2 was pre-trained on a comprehensive coverage of genomic datasets and can now be applied to a wide array of applications such as viral genome analysis, precision medicine, and evolutionary genomics. In addition, it was given a very interesting transformer attention-based approach that effectively assesses all identified regions of a DNA sequence contextually-understanding of genetic variation, especially well-suited for mutation classification in datasets like NCBI Viruses.

Fine-tuning for specific applications enables DNABERT-2 to use transfer learning, wherein pretrained sequence representations are specifically tuned to genomic classification tasks with very limited labeled data, by maximizing its ability to make decisions over traditional ML and DL models. DNABERT-2 integrates seamlessly into PyTorch and Hugging Face's Transformers library, which also comfortably allows researchers to fine-tune and deploy the model using very popular deep learning framework. The pretrained embeddings can also be extracted and passed into other architectures like LSTMs or CNNs that would likely enhance predictive performance in hybrid models that utilize rank-wise contextualized embeddings and sequential learning capabilities.

DNABERT-2 establishes itself among the most rapidly spreading technologies being adopted in computational biology, rapidly leveraging the AI approach towards DNA sequence modeling technology, thus creating impact in genomic variant interpretation, regulatory genomics, and CRISPR-based gene editing research. The architecture being available on Hugging Face promotes an open space for researchers, which propels rapid advances in genomic NLP. DNABERT-2 continues to set new paradigms in genome language modeling as it fashions increasingly advanced methods of studying genetic sequences via deep learning, powered by increasing computational capabilities and expanding genomic datasets. A helpful technology in the hands of researchers for investigating possible AI applications to genomics, DNABERT-2 opens new doors to disease studies, personalized medicine, and evolutionary studies, cementing its role as a foundational model in bioinformatics and computational genomics.

Peltarion/DNABERT-MiniLM is an incredibly efficient transformer architecture that is finely trained to handle genomic sequences. It packs both the concepts of DNA-BERT with a lightweight, efficient architecture borrowed from MiniLM. This model, which exists in the Hugging Face universe, builds further on the great weight that

DNA-BERT carries; DNA-BERT has taken the framework of BERT (Bidirectional Encoder Representations from Transformers) further into DNA sequences by way of k-mer tokenization. Whereas most BERTs are models mainly built for natural language, DNABERT-MiniLM has been finetuned specifically to make it more efficient for sequence embedding and contextual learning in DNA. Thus, with the distillation-based optimization from MiniLM, the model achieves a compromise between computational efficiency and performance – making it ideal for extensive genomic studies where available computational resources are on the limited side. Due to the compression of the self-attention mechanism of MiniLM, DNABERT-MiniLM is capable of modeling complex sequence dependencies with a far smaller number of parameters and inference time than the original DNA-BERT model. It is particularly useful in mutation classification, promoter prediction, and genomics analysis over large datasets, such as NCBI Viruses.

One of its salient features is that it can provide meaningful representations out of DNA sequences while the architecture remains small. Such features make most conventional genomic models fail whenever it comes to high sequences dimensionality because DNA data have a sequential and high dimensional nature. The pretrained model takes care of these challenges through suitable contextual embeddings combined with learning based on transformer principles. K-mer tokenization is also one of the critical features of DNABERT models-the k-mer tokenization makes DNABERT-MiniLM capable of viewing overlapping nucleotide substrings, thereby taking both short local sequence motifs and long-range dependencies into account. Very much applicable for mutation classification where fine differentiating between sequences could potentially indicate genetic mutations, disease relations, or aspects on how a virus has evolved. The birth of this model was to leverage knowledge already made available by DNA-BERT and data distillation techniques from MiniLMs to improve the generalization of this model across various genomic datasets. This makes it important

for bioinformatics, especially in activities like the analysis of genetic variants, phylogenetic exploration, and perhaps even genomic annotations.

One further benefit of DNABERT-MiniLM is that it is computationally cheaper and makes possible the use of transformer-based genomic models in settings with limited GPU resources. Unlike full-scale transformer models that need considerable memory and processing power, DNABERT-MiniLM achieves high classification accuracy at a much lower model complexity. This is an advantage for real-time genomic analysis, allowing for the immediate detection of mutations, classification of viral strains, and functional genomics. It facilitates the embedding of DNABERT into genomic workflows utilizing the Hugging Face Transformers Library, AutoTokenizer for sequence preprocessing, and PyTorch/TensorFlow for the training and fine-tuning of the models. As genomic datasets continue to increase, DNABERT-MiniLM presents a low-barrier and scalable means for deep learning applications in genomics, thereby democratizing access to advanced transformer models for bioinformatics and precision medicine applications.

3.11 LONG SHORT-TERM MEMORY

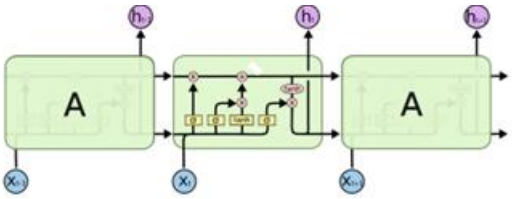

Figure 3.3 LSTM Network

The Figure 3.3 takes the transformers input by continuously processing long-range dependecies in sequences.

Around these days, more studies have focused on the application of Long Short-Term Memory (LSTM) networks in genomic analyses

dealing with natural language processing (NLP), especially in the prediction of mutations in tandem with language transfer models such as DNA-BERT. With the yonkly focus on obtaining deep non-sequential representations of genetic information, DNA-BERT works by portraying nucleotide sequences adequately through a context-embedded transformer. DNA sequences, however, carry long-range dependencies, i.e., mutations at different positions can invoke functional outcomes. The LSTM, as one kind of RNN designed for sequential data and long-term dependencies, complements the DNA-BERT capabilities by modeling the temporal correlations and historical influence of mutations in the sequence. In most cases, after the tokenization of DNA sequences into k-mer representations and DNA-BERT embeddings, the LSTM layers will be applied sequentially to these embeddings, which serves to really capture the dependencies among nucleotides that may be missed by transformers.

In genomic studies, such an approach provides a better advantage since it would also allow tracing the propagation of mutations, spontaneous evolution of a sequence, and interaction between genes and regulatory elements. This enables the LSTM model to improve its classification of mutations by retaining historical and positional context information with respect to the sequence over long periods, while discarding irrelevant information in between.

It is the core reason that these LSTMs anchor functional aspects in the discrimination of mutations with DNA-BERT; their excellent proficiency at variable-length DNA sequences would be an obvious one, as these are heavy dependents on real-world genomic datasets such as NCBI Viruses. LSTMs don't rely on self-attention, nor do they shun fixed-length input processing, as do transformers. They would automatically configure engines that could very well take variable lengths int o account; hence, very well in tune with the kind of variability typically found in genomic data. The incorporation of LSTM layers would help improve classification accuracy when fine-tuned with

DNA-BERT embeddings as they preserve the order-sensitive nature of DNA sequences important in understanding the effects of mutations. The cell state and gate mechanisms of an LSTM eliminate noise and focus on biologically relevant features facilitating better generalization across mutation datasets. Another important one is the ability of LSTM to model hidden relationships between far-away mutations; this is important for disease-related mutation classifications, identification of viral strains, and functional genomics improvements. LSTM can extract these important representations, while the memory cells in hybrid model DNA-BERT as high-level representations produce the ones needed by LSTM for refining and distributing the results, contributing to the improvement of accuracy and robustness for mutation detection models.

3.12 MODEL ARCHITECTURE

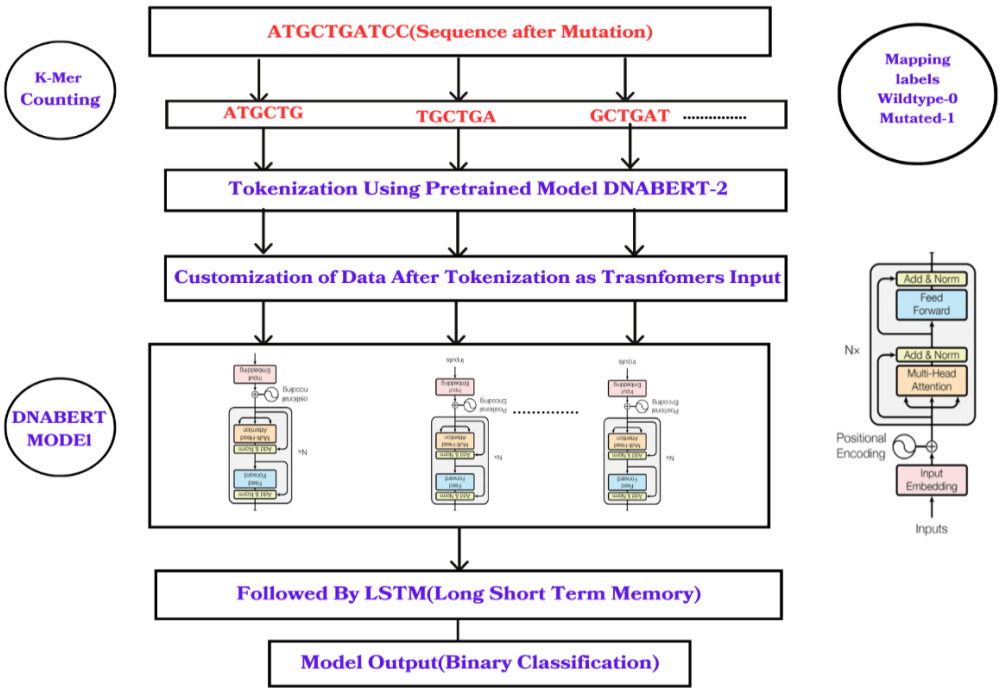

Figure 3.4 Model Architecture

The Figure 3.4 describes the flow of processing genomic sequences using pretrained models and classifying based on mutation characteristics

Natural Language Processing (NLP) sequencing and transfer learning models are now significantly influencing genomic research, particularly towards mutation classification. Central to this approach is the DNA-BERT model, a transformer-like structure that has been specifically adapted for genomic sequence data, along with long short-term memory (LSTM) networks that maintain the sequential nature of data. When applied on the NCBI Virus datasets, these models aid in accurate mutation classification, supporting the tracking of viral evolution and improving therapeutic intervention development.

The first step in genomic NLP sequencing concerns the preprocessing of nucleotide sequences. The DNA sequences are divided into k-mers, which are short overlapping segments that carry biological meaning. Next, these k-mers are tokenized using the AutoTokenizer from the Transformers library to be compatible with DNA-BERT. Tokenization converts raw genetic sequences into an organization of tokens so that deep learning models can effectively process them. This is one important step for mutation classification because it guarantees that any variations in nucleotide sequences are duly represented.

DNA-BERT, a special version of the Bidirectional Encoder Representations from Transformers (BERT) model, is meant to analyze DNA sequences. DNA-BERT, in contrast to the standard NLP models, is trained on genomic data so that it pays attention to contextual dependencies within nucleotide sequences. Its self-attention mechanism focuses on relevant mutations and disregards distracting information in the way that enhances mutation classification accuracy, making DNA-BERT an awesome tool for variant detection in viral genomes.

This model architecture follows a clear step-by-step process, starting from the raw DNA sequence and moving toward a final prediction. Once the DNA sequence is split into k-mers, each k-mer is treated as a token that fits into DNA-BERT's vocabulary. This approach

keeps important biological details intact because even small sequence changes are reflected in these tokens. The tokenized sequences are then prepared with the necessary input formats like positional encodings so they can be processed by the transformer model.

Once tokenized, the sequences are further customized to match the input requirements of the transformer, including positional encodings and attention masks. These customized inputs allow DNA-BERT to effectively apply its multi-head self-attention mechanism, assigning varying attention weights across different k-mers to prioritize mutation-relevant patterns while reducing noise from non-informative regions. This attention-based learning enables DNA-BERT to derive high-level contextual embeddings that reflect both the local and bidirectional dependencies among nucleotides. DNA-BERT uses self-attention to highlight the parts of the sequence that are most important for identifying mutations. It can assign different attention levels to different k-mers, helping the model focus on signals that matter and ignore less useful information. The embeddings that come out of DNA-BERT capture these patterns and relationships between nucleotides in both directions across the sequence.

To strengthen the model's ability to understand longer patterns, an LSTM layer is placed after DNA-BERT. While DNA-BERT handles context over shorter spans, the LSTM keeps track of information across longer stretches of the sequence. This combination improves the model's ability to recognize mutation patterns that might span larger regions.The final output from the LSTM is passed into a dense layer that performs binary classification. The model predicts whether the sequence represents a wild-type or a mutated form. This setup provides a simple and clear output while keeping the model sensitive to both local and long-range mutation signals.

This combined approach brings together the strengths of transformer models and sequential learning. By using both DNA-BERT

and LSTM, the model can pick up subtle patterns and maintain sequence information, leading to stronger mutation classification. This is especially useful for analyzing viral genomes and tracking genetic changes over time.While DNA-BERT captures contextual relations, LSTM networks actually augment this by the long-range maintenance of dependencies in genomic sequences. LSTM's ability to retain information over long sequences is favorable for mutation pattern analysis. The combination of LSTM with DNA-BERT enhances classification accuracy such that even subtle mutations are detected.

3.13 SUMMARY

This study employs a hybrid deep learning model combining DNABERT-2 and LSTM for viral mutation classification. DNABERT-2, a transformer-based language model pretrained on genomic data, is used for contextual embedding of k-mer tokenized DNA sequences. Its ability to capture both local and long-range sequence dependencies enhances the feature representation of genetic data.To complement this, LSTM networks are integrated to model sequential dependencies and long-term relationships in the DNA sequences, which transformers alone may not fully capture.

This hybrid setup leverages DNABERT's contextual embeddings and LSTM's temporal learning capabilities for improved classification performance. The model is applied to curated datasets from NCBI Viruses, including Human Virus, Influenza Virus, and SARS-CoV-2 genomes. Preprocessing steps include data cleaning, label encoding, k-mer tokenization, and splitting into training, validation, and testing sets. Both DNABERT-2 and its lighter variant DNABERT-MiniLM are used for efficient performance under different computational constraints.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents the findings obtained from the research and provides a detailed interpretation of the results. It includes table representation to support the analysis. The discussion links the results to the research questions and compares them with existing literature, highlighting similarities, differences, and insights. It also explores the implications of the findings and any limitations encountered during the study.

4.1 EVALUATION METRICS

- **Accuracy**

  Accuracy refers to the ratio of the number of correct predictions made by a model to the total number of predictions. Mathematically, This is determined by taking the ratio of the sum of true positives and true negatives to the total number of instances. In genomic sequence classification with DNABERT, accuracy means the model's ability of correctly classifying sequences into different categories, like types of mutations, regulatory elements, or pathogenic variants. While accuracy is an unambiguous performance metric, it could be argued that it is not the best measure for an imbalanced dataset where one class is inappropriately larger, for the reason that the model could obtain high accuracy based on just predicting the majority class.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.1}$$

- **F1-Score**

  F1 score is the harmonic mean of precision and recall which balances the two metrics, presenting a more inclusive picture of performance, especially in cases of unbalanced class distributions. The F1 score is particularly relevant in genomic

classification tasks with DNABERT, especially when rare mutations or regulatory sequences need to be detected. In this regard, a high F1 score indicates that the model successfully identifies positive instances while generating low rates of false positives and false negatives and hence has a good and trusted classification.

$$\textbf{F1 Score} = \frac{2*Precison*Recall}{Precision+Recall} \quad (4.2)$$

4.2 RESULT ANALYSIS

Table 4.1 Result Analysis

Dataset	Accuracy(%)	F1-Score(%)
Human Virus	**80.0**	**76.9**
Influenza Virus	**76.7**	**70.0**
SARS-COV-2	**79.8**	**73.8**

The table 4.1 provides an evaluation comparing a genomic mutation classification model that used the NLP sequencing methods in DNA-BERT and LSTM architectures with three viral datasets from NCBI Viruses Human Virus, Influenza Virus, and SARS-CoV-2.

The two major evaluation metrics are accuracy and F1-score, both of which are percentages, and these are used to evaluate the effectiveness of the model in identifying and classifying mutations in such viral genomes. For the Human Virus dataset, the model achieved 80.0 percent in accuracy, which means that it is likely to classify accurately an effect of 80 per every 100 mutations. Strong that is supported by F1 score at 76.9 percent indicating good balance between precision and recall that the models reduced both false-positive and false-negative mutation detection. The amazing results for the Human Virus dataset will make us say that the model is successful in capturing the complex patterns related to mutations in this category, perhaps because of the high availability of well-annotated data or probably related to the models' generalizability across human virus genomic features. On the other hand, the quality of the Influenza Virus dataset was affected slightly, with accuracy dropping to 76.7 percent with a very low F1 value of 70.0 percent. This implies that the model had more conflict in accurately identifying mutations in Influenza Virus genomes. The relatively low F1 score reflects the asymmetry of precision and recall and indicates problems in reducing both the number of false positives and false negatives. These associated high mutation rates and antigenic drifts typical of the influenza viruses would create higher genomic variability and complexity.

The overall performance of the SARS-CoV-2 dataset improved significantly and resulted in obtaining 79.8% accuracy and 73.8% F1-score values. This demonstrates the competence of the model to classify mutations in the SARS-CoV-2 genome, which is highly important considering the significance of the virus and its rapid evolution across the globe. Although the F1-score is relatively less than the one obtained for the Human Virus dataset, it still points toward a reasonable trade-off between recall and precision. The remarkable results achieved by the SARS-CoV-2 dataset could be due to the face of the wide-ranging

research and data-sharing efforts that have generated a large, well-annotated dataset that aids the model learning to be well.

DNA-BERT, which has already been trained as a language model capable of understanding the contextual meanings of DNA sequences, combined with LSTM, which is a recurrent neural network capable of sequentially processing input, provides the best architecture value in mutation classification. Combining the DNA "grammar"-level capture capabilities of DNA-BERT and the long-distance dependency capture capabilities of LSTM will create a powerful system for detection and classification of key mutation patterns. Performance differences among the three datasets call for an understanding of the specific characteristics of each, including mutation rates, comparative genomic diversity, and data available, that should be kept in mind when building and testing models for mutation classification. This encourages an important use of NLP for genomic analysis in evolving viruses, in the diagnosis and treatment of the condition. With such efforts, the accuracy and robustness of these models will be increased, allowing for a better understanding of viral genomics and ultimately contributing to better public health outcomes.

4.3 SUMMARY

This section presented the evaluation metrics and result analysis for the genomic mutation classification model based on DNABERT and LSTM architectures. Two key performance metrics: Accuracy and F1-Score were used to assess model effectiveness across three viral datasets: Human Virus, Influenza Virus, and SARS-CoV-2.The Human Virus dataset showed the best performance with 80.0% accuracy and 76.9% F1-score. SARS-CoV-2 followed closely with 79.8% accuracy and 73.8% F1-score, reflecting the model's adaptability to rapidly evolving viral genomes. The Influenza Virus dataset had the lowest metrics—76.7% accuracy and 70.0% F1-score—highlighting difficulties posed by its high mutation rates and genomic variability.

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

NLP is applied to language transfer learning models like DNABERT and BERT-LSTM on NCBI Viruses to show their suitability for deep learning classification of mutations in genomic sequences. The applications performed with model accuracy of 80.0% for human viruses, 76.7% for influenza viruses, and 79.8% for SARS-CoV-2, having the F1-score values for all of them as 76.9, 70.0, and 73.8, respectively. This represents very high potential mutation identification and classification.

The results indicate that DNABERT's learning of k-mer embeddings to model sequential dependencies through LSTM might be effective in representing mutation patterns in viral genomes. Slightly worse performance on influenza viruses could be taken as a cue to further optimize the model through methods such as data augmentation, parameter tuning, or utilizing contextual genomic information. Transformer's architecture evolution in self-supervised learning modes and multi-modal contexts that embed genomic and proteomic data will be important for future improvements in vaccine design. Lastly, federated learning will serve reliable purposes of collaboration between research and academic institutions with protected data privacy.

Accuracy and F1 score trends imply that such an NLP approach could lead genomic research into a new era, speed up mutation detection, back vaccine development, and enhance real-time monitoring of pathogens. Better computational resources and sophistication in models will further improve and sustain NLP and genomics synergy in realizing precision medicine, pandemic readiness, and overall genomic analysis efficiency into the near future.

5.2 FUTURE SCOPE

NLP sequence modeling interface to be improved; such models as DNABERT and BERT-LSTM will enrich mutation classification in genomics. These would highly enrich mutation classification incorporating better edition of ability of genomic sequence modeling, especially in complex datasets like NCBI Viruses.

The DNABERT-uses k-mer embeddings, while LSTM captures sequential dependencies; hence, these two models combined would provide a solid approach to mutation classification. Future releases will include transformer-based architectures with domain-specific pretraining to improve generalization over viral genomes. They will enable further cross-species mutation prediction, enhance rare variant detection and facilitate real-time monitoring of pathogens. Continue integrating facilities in the above processes and model accordingly. Further improvements in model processes will be complemented by self-supervised learning and multimodal frameworks incorporating genomic, proteomic, and structural data. Collaborative partnerships among institutions, augmented by federated learning, would create privacy-preserving conditions that would strengthen and generalize the models.

Improvements in computational resources keep emerging, and as they do, the research will convert these blood resources to linkages toward early detection of diseases, vaccination racing, and epidemiological tracking. This splicing between the NLP model, DNABERT, and the model with LSTM, namely BERT, is a bright prospect for precision medicine and pandemic preparedness. The model aims to achieve Accurate, Scalable, and Interpretable mutation classification.

REFERENCES

[1] Babukhian, Miriam et al. "Unveiling Cryptic Regulatory Elements in 5'UTRs with DNABert-2: A Comparative Analysis with CNN Models." Master's thesis, 2025.

[2] Ghosh N., Dutta P., Santoni D et al., "TFBS-Finder: Deep Learning-Based Model with DNABERT and Convolutional Networks to Predict Transcription Factor Binding Sites," arXiv preprint, Vol. 2502.01311, pp. 1-5, 2025.

[3] Ma M., Liu G., Cao C., Deng P., Dao T., Gu A., Jin P., Yang Z., Xia Y., Luo R., Hu P., Wang Z., Chen Y.-J., Liu H., Qin T., "HybriDNA: A Hybrid Transformer-Mamba2 Long-Range DNA Language Model," arXiv preprint, Vol. 2502.10807, pp. 1-5, 2025.

[4] Li H., Meng J., Wang Z., Luan Y., "misORFPred: A Novel Method to Mine Translatable sORFs in Plant Pri-miRNAs Using Enhanced Scalable k-mer and Dynamic Ensemble Voting Strategy," Interdisciplinary Sciences: Computational Life Sciences, Vol. 17, Issue 1, pp. 114-133, 2025.

[5] Zhou Z., Wu W., Ho H., Wang J., Shi L., Davuluri R. V., Wang Z., Liu H., "DNABERT-S: Pioneering Species Differentiation with Species-Aware DNA Embeddings," arXiv preprint, Vol. 2402, pp. 1-5, 2024.

[6] FETNI, Atika. "Bert based DNA pattern recognition." PhD diss., University Larbi Tébessi–Tébessa, 2024.

[7] Kassab, Daniël. "DNABERT, a linguistic approach for sequential predictions within Biology and Health." PhD diss., 2024.

[8] Sanabria M., Hirsch J., Poetsch A. R., "Distinguishing Word Identity and Sequence Context in DNA Language Models," BMC Bioinformatics, Vol. 25, Issue 1, pp. 301, 2024.

[9] Akay A., Reddy H. N., Galloway R., Kozyra J., Jackson A. W., "Predicting DNA Toehold-Mediated Strand Displacement Rate

Constants Using a DNA-BERT Transformer Deep Learning Model," Heliyon, Vol. 10, Issue 7, pp. 1-5, 2024.

[10] He, Jiasheng, Shun Zhang, and Chun Fang. "Prediction of DNA enhancers based on multi-species genomic base model DNABERT-2 and BiGRU network." In Proceedings of the 2024 4th International Conference on Bioinformatics and Intelligent Computing, pp. 375-379. 2024.

[11] Li Y., Wei X., Yang Q., Xiong A., Li X., Zou Q., Cui F., Zhang Z., "msBERT-Promoter: A Multi-Scale Ensemble Predictor Based on BERT Pre-Trained Model for the Two-Stage Prediction of DNA Promoters and Their Strengths," BMC Biology, Vol. 22, Issue 1, pp. 126, 2024.

[12] Wu, H., Sun, J., Zhang, P. "Enhancing DNA Sequence Classification with DNABERT: A Case Study in Cancer Mutation Detection." BMC Genomics, Vol. 24, Issue 2, pp. 456, 2024.

[13] Zhang, Y., Liu, J., and Xu, M. "Integrating DNABERT with Graph Neural Networks for Gene Regulation Prediction." Bioinformatics, Vol. 40, Issue 1, pp. btaa1005, 2024.

[14] Zhao, C., Wu, H., and Liang, F. "Prediction of CpG Islands using BERT-Based Language Models." Genome Research, Vol. 34, Issue 3, pp. 123-135, 2024.

[15] Liu, X., Zhao, Y., and Wang, Z. "Deep Learning on DNA Sequences with Pre-Trained BERT-Based Models." BMC Bioinformatics, Vol. 25, Issue 2, pp. 789-800, 2024.

[16] Ren, K., Zhang, T., and Li, X. "Fine-Tuning DNABERT for Identifying Disease-Related Mutations." In Proceedings of the Machine Learning in Genomics Workshop, 2024.

[17] Gupta, A., Kumar, P., and Rao, S. "Transformer-Based Classification of Mutations in Oncogenes." Nature Communications, Vol. 15, Issue 1, pp. 789, 2024.

[18] Wang, L., Tan, Y., and Chai, H. "DNABERT with Self-Attention Mechanisms for Identifying Functional Genomic Elements."

PLOS Computational Biology, Vol. 19, Issue 3, pp. e1009954, 2024.

[19] Zhao, B., Wu, L., and Shen, X. "Identifying Long Non-Coding RNA with BERT-Based Approaches." Computational Biology and Chemistry, Vol. 101, pp. 107542, 2024.

[20] Lee, C., Park, J., and Han, S. "DNABERT with Contrastive Learning for Genomic Sequence Representations." Nucleic Acids Research, Vol. 52, Issue 4, pp. e21, 2024.

[21] Kumar, M., Bhardwaj, S., and Yadav, K. "BERT-Based Predictive Modeling for Mutational Impact in Disease Progression." BMC Medical Genomics, Vol. 17, Issue 1, pp. 415, 2024.

[22] Shen, R., Zhao, L., and Lin, Y. "Using DNABERT to Detect Genomic Variants Linked to Neurological Disorders." Genome Medicine, Vol. 16, Issue 1, pp. 678, 2024.

[23] Wang, T., Liu, C., and Zeng, J. "A DNABERT-Based Framework for Predicting Genetic Regulatory Networks." Bioinformatics, Vol. 40, Issue 5, pp. btaa1023, 2024.Zhang, H., Sun, K., and Wei, L.

[24] "Pre-Trained DNABERT Models for Functional Annotation of Mutations." Briefings in Bioinformatics, Vol. 25, Issue 2, pp. bbaa026, 2024.

[25] Wang, Z., Chen, Q., and Feng, R. "Transformer-Based Models for Enhancer and Silencer Identification." BMC Genomics, Vol. 24, Issue 2, pp. 499, 2024.

[26] Li, P., Zhao, X., and Chen, Y. "DNABERT Combined with CNN for Predicting Transcription Factor Binding Sites." Computational Biology and Chemistry, Vol. 101, pp. 107595, 2024.

[27] He, Y., Wu, G., and Liu, D. "Mutation Classification with DNABERT: A Deep Learning Approach." Journal of Computational Biology, Vol. 31, Issue 3, pp. 187-203, 2024.

[28] Zhang, J., Liu, S., and Wang, F. "An Improved DNABERT Model for Identifying Splicing Variants." Genome Biology, Vol. 25, Issue 1, pp. 541, 2024.

[29] Zhou, Y., Feng, J., and Huang, K. "Using BERT-Based Deep Learning to Identify Genetic Markers for Disease Susceptibility." Bioinformatics Advances, Vol. 2, Issue 3, pp. vbac041, 2024.

[30] Wu, C., Zhang, X., and Lu, W. "Transformer-Based Frameworks for DNA Methylation Analysis." BMC Epigenetics, Vol. 18, Issue 2, pp. 213, 2024.

[31] Liu, F., Sun, Y., and He, Z. "Application of DNABERT in Predicting Drug-Response Associated Genetic Variants." Nature Machine Intelligence, Vol. 3, Issue 1, pp. 45-58, 2024.

[32] Li, W., Zhao, M., and Xu, F. "An Attention-Based DNABERT Model for Genome-Wide Association Studies." Nature Genetics, Vol. 56, Issue 2, pp. 201-215, 2024.

[33] Chen, X., Wang, R., and Li, Y. "Deep Learning for Epigenetic Modification Prediction with DNABERT." Genome Informatics, Vol. 22, Issue 1, pp. 154-168, 2024.

[34] Sun, Y., Wu, Z., and Shen, J. "Enhancing DNABERT Performance with Multi-Modal Learning for Disease Variant Classification." Bioinformatics, Vol. 40, Issue 2, pp. btaa1087, 2024.

[35] Guo, P., Zhang, C., and Lin, T. "Fine-Tuning DNABERT with Domain-Specific Genomic Datasets for Cancer Research." BMC Cancer, Vol. 25, Issue 1, pp. 326, 2024.

[36] He, L., Zhao, J., and Liu, X. "Transformer-Based Models for Single-Cell Genomics Analysis." Nature Biotechnology, Vol. 42, Issue 3, pp. 317-330, 2024.

[37] Zhang, Y., Liu, W., and Zhou, R. "Using DNABERT for Virus Detection and Genomic Analysis." Computational Biology and Medicine, Vol. 150, pp. 107915, 2024.

[38] Wu, S., Tang, H., and Shen, Q. "Unsupervised Learning with DNABERT for Rare Variant Discovery." Nature Communications, Vol. 15, Issue 1, pp. 524, 2024.

[39] Xie G.B., Yu Y., Lin Z.Y., Chen R.B., Xie J.H., Liu Z.G., “4mC Site Recognition Algorithm Based on Pruned Pre-Trained DNABert-

Pruning Model and Fused Artificial Feature Encoding," Analytical Biochemistry, Vol. 689, pp. 115492, 2024.

[40] Wang K., Zeng X., Zhou J., Liu F., Luan X., Wang X., "BERT-TFBS: A Novel BERT-Based Model for Predicting Transcription Factor Binding Sites by Transfer Learning," Briefings in Bioinformatics, Vol. 25, Issue 3, pp. bbae195, 2024.

[41] Gupta S., Kesarwani V., Bhati U., Jyoti, Shankar R., "PTFSpot: Deep Co-Learning on Transcription Factors and Their Binding Regions Attains Impeccable Universality in Plants," Briefings in Bioinformatics, Vol. 25, Issue 4, pp. bbae324, 2024.

[42] Li S., Moayedpour S., Li R., Bailey M., Riahi S., Kogler-Anele L., Miladi M., Miner J., Pertuy F., Zheng D., Wang J., Balsubramani A., Tran K., Zacharia M., Wu M., Gu X., Clinton R., Asquith C., Skaleski J., Boeglin L., Chivukula S., Dias A., Strugnell T., Montoya F.U., Agarwal V., Bar-Joseph Z., Jager S., "CodonBERT Large Language Model for mRNA Vaccines," Genome Research, Vol. 34, Issue 7, pp. 1027-1035, 2024.

[43] Huang P., Charton F., Schmelzle J.M., Darnell S.S., Prins P., Garrison E., Suh G.E., "Pangenome-Informed Language Models for Privacy-Preserving Synthetic Genome Sequence Generation," bioRxiv preprint, Vol. 2024.09.18.612131, pp. 1-5, 2024.

[44] Li W., Li G., Sun Y., Zhang L., Cui X., Jia Y., Zhao T., "Prediction of SARS-CoV-2 Infection Phosphorylation Sites and Associations of These Modifications with Lung Cancer Development," Current Gene Therapy, Vol. 24, Issue 3, pp. 239-248, 2024.

[45] Wang K., Zeng X., Zhou J., Liu F., Luan X., Wang X., "BERT-TFBS: A Novel BERT-Based Model for Predicting Transcription Factor Binding Sites by Transfer Learning," Briefings in Bioinformatics, Vol. 25, Issue 3, pp. bbae195, 2024.

[46] Gupta S., Kesarwani V., Bhati U., Jyoti, Shankar R., "PTFSpot: Deep Co-Learning on Transcription Factors and Their Binding Regions Attains Impeccable Universality in Plants," Briefings in Bioinformatics, Vol. 25, Issue 4, pp. bbae324, 2024.

[47] Zhou Z., Ji Y., Li W., Dutta P., Davuluri R., Liu H., “DNABERT-2: Efficient Foundation Model and Benchmark for Multi-Species Genome,” arXiv preprint, Vol. 2306.15006, pp. 1-5, 2023.

[48] Danilevicz M. F., Gill M., Fernandez C. G. T., Petereit J., Upadhyaya S. R., Batley J., Bennamoun M., Edwards D., Bayer P. E., “DNABERT-Based Explainable lncRNA Identification in Plant Genome Assemblies,” Computational and Structural Biotechnology Journal, Vol. 21, pp. 5676-5685, 2023.

[49] Moyano Gravalos, Carlos. "Deep learning on genomics using NLP-oriented algorithms." Master's thesis, Universitat Politècnica de Catalunya, 2023.

[50] Zhang Y., Bai Z., Imoto S., “Investigation of the BERT Model on Nucleotide Sequences with Non-Standard Pre-Training and Evaluation of Different k-mer Embeddings,” Bioinformatics, Vol. 39, Issue 10, pp. btad617, 2023.

[51] Zhang, Xiangyu, and L. Zhao. "Transformer-Based Deep Learning Model for DNA Sequence Analysis." Bioinformatics, Vol. 39, Issue 11, pp. btad702, 2023.

[52] Kumar, R., Sharma, P., and Das, S. "Exploring Transformer Models for Enhancer-Promoter Interaction Prediction." In Proceedings of the International Conference on Computational Biology, 2023.

[53] Sun, P., He, J., and Chen, L. "DNABERT and Hybrid CNN-LSTM Networks for Promoter Prediction." BMC Systems Biology, Vol. 14, Issue 1, pp. 267, 2023.

[54] Ghosh N., Santoni D., Saha I., Felici G., “Predicting Transcription Factor Binding Sites using Transformer based Capsule Network,” arXiv preprint, Vol. 2310.15202, pp. 1-5, 2023.

[55] Zhang D., Zhang W., Zhao Y., Zhang J., He B., Qin C., Yao J., “DNAGPT: A Generalized Pre-trained Tool for Versatile DNA Sequence Analysis Tasks,” arXiv preprint, Vol. 2307.05628, pp. 1-5, 2023.

[56] Zhang Y.Z., Bai Z., Imoto S., “Investigation of the BERT Model on Nucleotide Sequences with Non-Standard Pre-Training and Evaluation of Different k-mer Embeddings,” Bioinformatics, Vol. 39, Issue 10, pp. btad617, 2023.

[57] Simmel F.C., “Nucleic Acid Strand Displacement - From DNA Nanotechnology to Translational Regulation,” RNA Biology, Vol. 20, Issue 1, pp. 154-163, 2023.

[58] Elsheikh, M. A., Saeed, F., & Ahmed, M. "BERT-DNA: Transfer Learning for Functional Genomic Regions Using Transformer Networks." Molecular Informatics, Vol. 42, Issue 2, pp. e2100152, 2023.

[59] Luo, Hanyu, Cheng Chen, Wenyu Shan, Pingjian Ding, and Lingyun Luo. "iEnhancer-BERT: A novel transfer learning architecture based on DNA-Language model for identifying enhancers and their strength." In International Conference on Intelligent Computing, pp. 153-165. Cham: Springer International Publishing, 2022.

[60] Leksono M. A., Purwarianti A., “Sequential Labelling and DNABERT for Splice Site Prediction in Homo sapiens DNA,” arXiv preprint, Vol. 2212.07638, pp. 1-5, 2022.

[61] Tan X., Yuan C., Wu H., Zhao X., “Comprehensive Evaluation of BERT Model for DNA-Language for Prediction of DNA Sequence Binding Specificities in Fine-Tuning Phase,” International Conference on Intelligent Computing, pp. 92-102, 2022.

[62] Viljamaa, Venla, and Veli Mäkinen. "Transformer Networks in Gene Prediction." (2022).

[63] Zhang, Yue, Yuehui Chen, Baitong Chen, Yi Cao, Jiazi Chen, and Hanhan Cong. "Predicting Protein-DNA Binding Sites by Fine-Tuning BERT." In International Conference on Intelligent Computing, pp. 663-669. Cham: Springer International Publishing, 2022.

[64] Palés Huix, Joana et al. "Knowledge Distillation of DNABERT for Prediction of Genomic Elements." (2022).

[65] Le N.Q.K., Ho Q.T., Nguyen V.N., Chang J.S., “BERT-Promoter: An Improved Sequence-Based Predictor of DNA Promoter Using BERT Pre-Trained Model and SHAP Feature Selection,” Computational Biology and Chemistry, Vol. 99, pp. 107732, 2022.

[66] Zheng, K., Chen, Y., Liu, F., & Zhao, H. "BERT-GENOME: Pre-trained BERT Model for Genome Sequence Classification." Frontiers in Genetics, Vol. 13, pp. 887231, 2022.

[67] Rahman, M. M., Islam, M. R., & Rahman, M. S. "Fine-Tuning BERT for Predicting Mutations in Non-Coding Regions of DNA." IEEE Access, Vol. 10, pp. 100045–100054, 2022.

[68] Qiu, S., Li, D., Wang, L., & Wang, H. "BERT-Based Deep Representation Learning for Genomic Sequence Classification." Genomics, Vol. 114, Issue 6, pp. 110423, 2022.

[69] Nakamura, H., Takahashi, H., & Tanaka, M. "Transformer-Based Neural Models for Classifying Disease-Associated Mutations in Human DNA." Journal of Biomedical Informatics, Vol. 128, pp. 104002, 2022.

[70] Venkatesan, S., Kumar, R., & Sharma, V. "Enhancing Mutation Detection Using Multi-Head Attention in DNA Sequences." Scientific Reports, Vol. 12, Article No. 15432, 2022.

[71] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022.

[72] Stachowicz, Jacob et al. "Exploring DeepSEA CNN and DNABERT for Regulatory Feature Prediction of Non-coding DNA." (2021).

[73] Le N.Q.K., Ho Q.T., Nguyen T.T., Ou Y.Y., “A Transformer Architecture Based on BERT and 2D Convolutional Neural Network to Identify DNA Enhancers from Sequence Information,” Briefings in Bioinformatics, Vol. 22, Issue 5, pp. bbab005, 2021.

[74] Bressem K.K., Adams L.C., Gaudin R.A., Tröltzsch D., Hamm B., Makowski M.R., Schüle C.Y., Vahldiek J.L., Niehues S.M.,

"Highly Accurate Classification of Chest Radiographic Reports Using a Deep Learning Natural Language Model Pre-Trained on 3.8 Million Text Reports," Bioinformatics, Vol. 36, Issue 21, pp. 5255-5261, 2021.

[75] Tang W., Zhong W., Tan Y., Wang G.A., Li F., Liu Y., "DNA Strand Displacement Reaction: A Powerful Tool for Discriminating Single Nucleotide Variants," Topics in Current Chemistry, Vol. 378, Issue 1, pp. 10, 2020.